1-7-2016

Analyses of amber mutants of Salmonella phage SPN3US

Andrea Denisse Benítez Quintana
adb8932@rit.edu

R·I·T

Analyses of amber mutants of *Salmonella* phage SPN3US

by

Andrea Denisse Benítez Quintana

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science in Bioinformatics

Bioinformatics Program

College of Science

Rochester Institute of Technology

Rochester, NY

January 7, 2016

Rochester Institute of Technology
Thomas H. Gosnell School of Life Sciences
Bioinformatics Program

To: Head, Thomas H. Gosnell School of Life Sciences

The undersigned state that ____Andrea Denisse Benítez Quintana_____, a candidate for the Master of Science degree in Bioinformatics, has submitted his/her thesis and has satisfactorily defended it.

This completes the requirements for the Master of Science degree in Bioinformatics at Rochester Institute of Technology.

Thesis committee members:

Name	Date
(Thesis Advisor) Julie Thomas, Ph.D.	Jan 11, 2016
Michael Osier, Ph.D.	Jan 8, 2016
André Hudson, Ph.D.	Jan 11, 2016

Michael V. Osier, Ph.D.
Graduate Director of Bioinformatics

475-4392 (voice)
mvoscl@rit.edu

Contents

1. ABSTRACT

The giant *Salmonella enterica* phage SPN3US belongs to the ϕKZ-related phages which are amongst the most complex virions of prokaryotic viruses, even more complex than most eukaryotic viruses. SPN3US for example assigns 46% of its 240kbp genome to code for virion proteins. We have limited knowledge about these giant phages as they encode extremely diverged genes, from other phages and even to one another. The ϕKZ-related phages possess large genomes and correspondingly large and complex capsids, two sets of fragmented RNAP β and β'-like proteins and they also exhibit an unusual replication having a DNA polymerase split into two polypeptides.

The complete genome of phage SPN3US was previously sequenced by Lee et al. (2011). As is typical for a newly sequenced phage, 80% of its ORFs were annotated as hypothetical. To address this problem of functionally unassigned genes, a novel strategy for phages is to target the essential genes and work to determine their functions using genetic approaches. In this work amber mutants of a giant phage were sequenced for the first time. Once the amber mutation sites were determined the sequence similarity of those genes in other ϕKZ-related phages was analyzed. This has enabled us to identify fourteen essential genes in SPN3US, ten of which code for the components of the phage virion. Previously we knew the function of three genes, now we can infer the function of four more genes. The mutants encoding these essential genes will now form the basis of further studies to determine specific functions for these novel genes. Ten of these genes are well conserved among the ϕKZ-related phages being classified as core genes for this subfamily.

These studies have demonstrated that SPN3US is an excellent model system to study the ϕKZ-related phages.

2. ACKNOWLEDGEMENTS

Rochester Institute of Technology

- Dr. Julie Thomas
- Dr. Michael Osier
- Dr. Andre Hudson

University of Rochester Genomics Research Center

- John M. Ashton, Ph.D.
- Michelle Zanche, MS
- Jason Myers, MS

University of Maryland – School of Medicine

- Dr. Lindsay Black
- Elizabeth Aguilera
- Assitan Coulibaly

University of Texas Health Science Center in San Antonia (UTHSCSA)

- Dr. Susan T. Weintraub
- Kevin Hakala

A.D.B.Q was financially supported by Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación (SENESCYT - Ecuador)

3. LIST OF ABBREVIATIONS AND ACRONYMS

AA	Amino acid
ATP	Adenosine triphosphate
CDS	Conserved Domain Search
cryoEM	Cryo-electron microscopy
csv	Comma separated values
dsDNA	Double-stranded deoxyribonucleic acid
EDTA	Ethylenediaminetetraacetic acid
EM	Electron microscopy
FDA	Food and Drug Administration
gp	Gene product
GRAS	Generally recognized as safe
HA	Hydroxylamine
Kbp	Kilobase pairs
LB	Luria Broth
LB+N	Luria Nutrient Broth
NCBI	National Center for Biotechnology Information
nt	Nucleotide
nvRNAP	Non-virion ribonucleic acid polymerase
ORF	Open reading frame
PCR	Polymerase Chain Reaction
PFU	Plaque forming units
RNAP	Ribonucleic acid polymerase
SNP	Single nucleotide polymorphism
ssRNA	Single-stranded ribonucleic acid
sup^{+}	Suppressor
sup^{-}	Non-suppressor
TEM	Transmission electron microscopy
tRNA	Transfer ribonucleic acid
vRNAP	Virion ribonucleic acid polymerase

4. INTRODUCTION

4.1 Bacteriophages: history and structure

Bacterial viruses infect bacteria and use them as hosts in order to replicate. They were discovered independently by Frederick Twort (1915) in Great Britain and Felix d'Hérelle (1917) in France. Twort defined them as agents that were capable of killing bacteria but it was D'Hérelle who first described and named them bacteriophages (Greek for eaters of bacteria) or phages. D'Hérelle developed a test called the plaque assay with which individual clearing areas, plaques, could be seen in the bacterial lawn where the bacterial cells have been lysed (Anderson et al. 2011).

The genetic material of phages can be either single-stranded or double-stranded DNA or RNA, and may have notable length variations, ranging from the ~3300 nucleotides from the ssRNA viruses of *Escherichia coli* to the 497,531 base pairs from phage G that infects *Bacillus megaterium* (Hatfull and Hendrix 2011). Phages possess two main ways of infecting their host and propagating: lytic life cycle and lysogenic life cycle. Once the host is infected a phage can either go into the lytic life cycle in which the phage replicates inside the host and kills the cell by lysis, or into the lysogenic life cycle in which the phage genome is replicated benignly with the host chromosome for many generations. Temperate phages can reproduce with lysogenic cycle and may be induced into lytic cycle (Haq et al. 2012). Phage structures vary from tailed, polyhedral, filamentous, or pleomorphic; but the majority that have been isolated are tailed phages which are the most diversified phage group and belong to the order *Caudovirales* (Ackermann & Prangishvili, 2012).

4.1.1 *Caudovirales:* tailed phages

Tailed phages have a head or capsid based on icosahedral symmetry, of these 15% have prolate or elongated heads, and all of them have double stranded DNA genome (Ackermann & Prangishvili, 2012). The capsid contains the genome which is packaged by ATP-driven molecular motors that generally have two highly conserved elements: the terminase and the portal that contains a channel enabling the entrance and exit of the DNA (Black and Thomas 2012). They possess tail fibers, that help them attach to the bacteria cell, and the tail to deliver the genome into the cell and infect the bacteria (Ackermann, 1998).

The *Caudovirales* has three families characterized by different type of tails: *Myoviridae*, contractile tails*; Siphoviridae*, long and noncontractile tails; and *Podoviridae*, short tails (Fig 1).

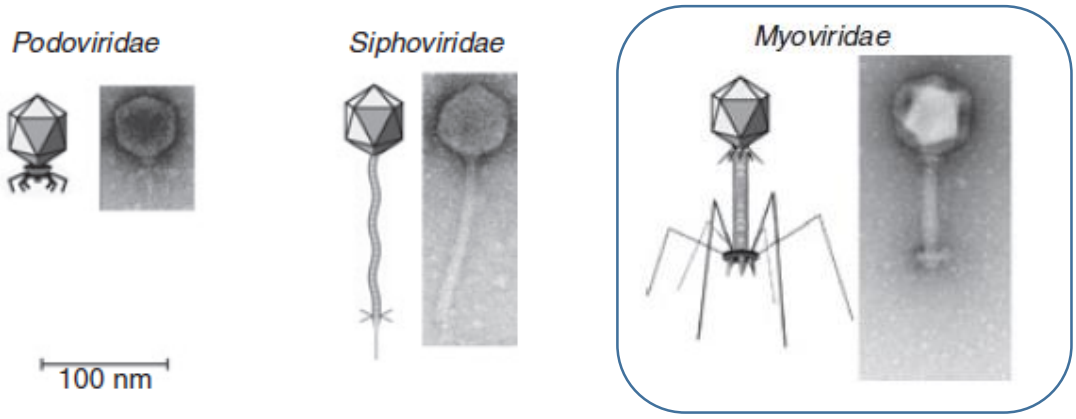

Figure 1: The tailed phages. The majority that have been isolated belong to the *Myoviridae* family, highlighted in blue (Adapted from Harper and Enright 2011).

The NCBI phage genome database contains 2048 complete sequenced genomes of dsDNA viruses (no RNA stage) and the majority (1427) belong to the order *Caudovirales* (www.ncbi.nlm.nih.gov, Retrieved September, 2015). This database has shown an exponential increment since 2000 denoting an increasing interest in phage research (Fig 2).

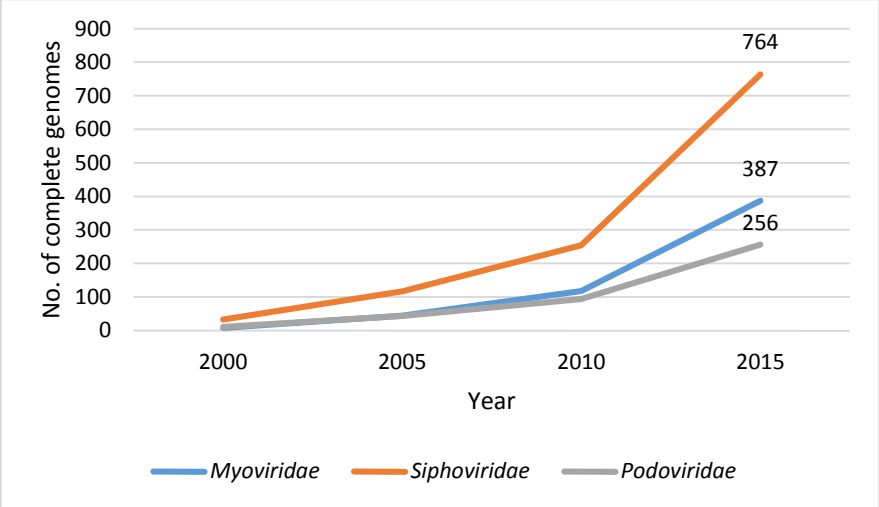

Figure 2: Plot showing the numbers of individually sequenced genomes in each tailed phage family in the last fifteen years. The number of complete genome sequences was obtained from Viral Genomes in September 2015.

4.1.2 Abundance and diversity of phages

The abundance of phages found on earth is astonishing, the size of this population has been estimated to be 10^{30}-10^{31} phages globally (Hendrix et al. 1999). Phages constitute the most numerous group of all viruses (Ackermann & Prangishvili, 2012) as they can be found in every ecosystem where bacteria thrive: surface seawater, marine sediment, freshwater, soil, and even in extreme environments including hot springs, hypersaline lakes and solar salterns (Sano et al. 2004) and they outnumber their hosts in a ratio of 10:1 (Hendrix et al. 1999).

The origin of phages could be comparable to the origin of their bacterial hosts making them a very old population, originating more than three billion years ago. Most tailed phages share a common gene pool due to a common ancestry, no matter the environment of the host (Hendrix et al. 1999; Hatfull and Hendrix 2011); however, they also have a high degree of genetic diversity (Hatfull 2008). In a metagenomics study performed by Breitbart et al. (2002) approximately 65% of the sequences obtained from the uncultured marine viral communities returned no homologues when compared with nucleotide databases, which suggests that the vast majority of viral diversity has not been characterized (Breitbart et al. 2002). Even though phages are the most abundant and diverse biological entities on the planet only a small percent has been studied in detail, "phages might indeed represent the largest unexplored reservoir of sequence information in the biosphere" (Chibani-Chennoufi et al. 2004, p. 3680). About 80% of their encoded genes are not related to known proteins and their function is unknown (Hatfull and Hendrix 2011).

4.2 Relevance of phages to humans

After discovering phages Felix d'Hérelle realized they could be used as tools against infectious diseases caused by bacteria such as cholera and bubonic plague marking the beginning of phage therapy (Nelson et al. 2009). This first age of phage therapy ended in the United States and the Western European countries with the development of chemical antibiotics after World War II, phages continued to be used therapeutically in the former Soviet Union and Eastern Europe (Fig 3) (Sulakvelidze et al. 2001; Haq et al. 2012). One of the main reasons for the loss of interest in phage therapy was the poor understanding of phage biology. For instance, the concept of host-range (the specificity of a phage for a particular host taxon, e.g., phages that infect pseudomonads never infect bacilli) was not understood and so for a time inappropriate phages for therapy were selected (Haq et al. 2012).

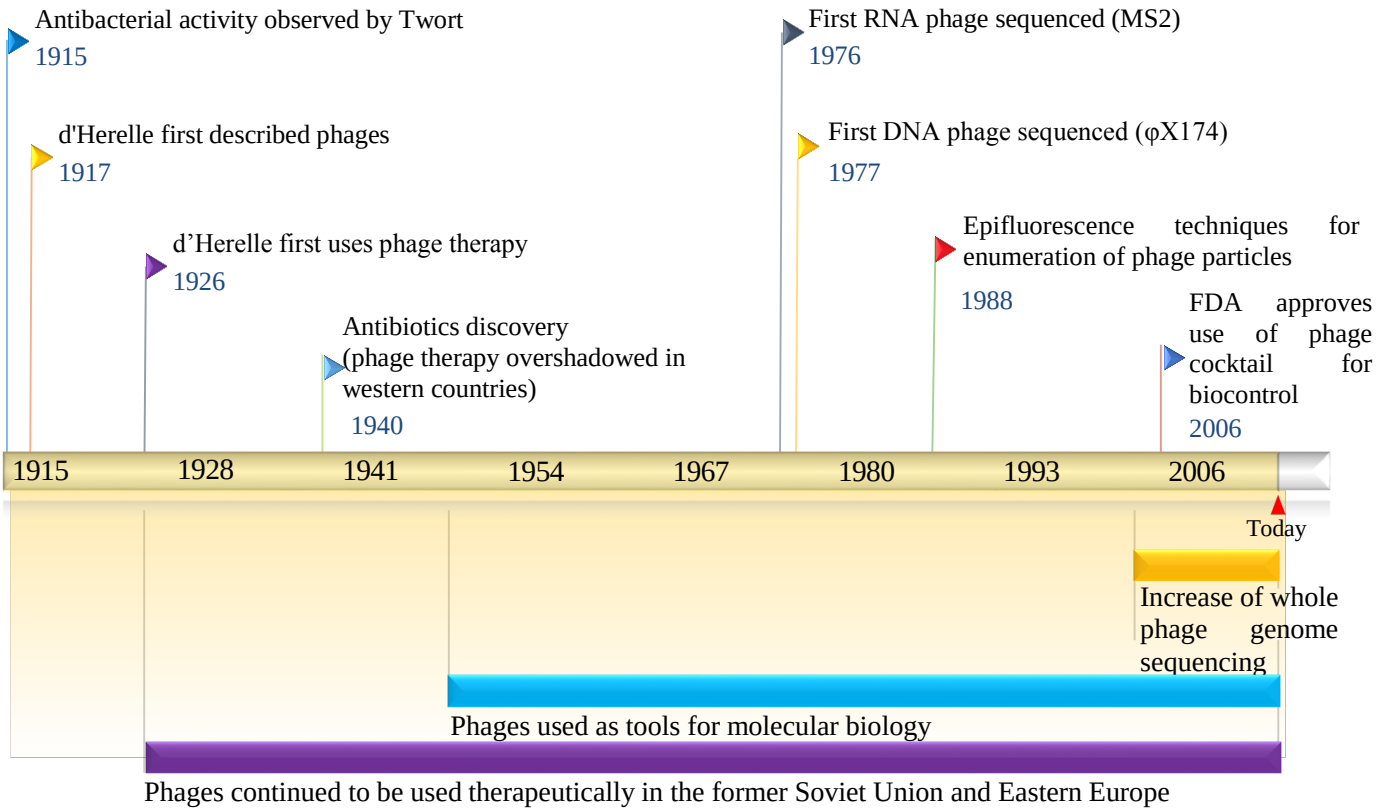

Figure 3: Timeline displaying the major milestones in phage history and its applications.

Later on in the mid-20th century phages played an important role in the development of molecular biology, being relatively simple biological systems they helped understand the basic molecular mechanisms of gene replication and gene structure, and were considered "the cradle of molecular biology" (Chibani-Chennoufi et al. 2004, p. 3677). Their simplicity of isolation and relatively small size were the reasons why phages were chosen to be the first organisms to be completely sequenced. In the 1980's the great abundance of phages was revealed thanks to epifluorescence techniques used to enumerate phage particles in environmental samples (Fig 3). The interest in the great abundance and genetic diversity of phages and advances in DNA sequencing technology led to a new age of phage research, focusing on the important roles they have in evolution, ecology, bioremediation, biocontrol and phage therapy, which will be covered in more depth in the next section (Mann 2005; Hatfull 2008).

Evolutionarily phages are a key factor for horizontal gene exchange between different bacterial species, they also constitute the main reason for the evolution of their hosts obliged to develop phage resistance (Krupovic et al. 2011) and as such are key in maintaining bacterial genetic diversity (Mann 2005). Phages themselves possess mosaic genome architectures, suggesting horizontal genetic exchange among them (Hatfull and Hendrix 2011), and is usually found in phages from related bacterial species, although some examples have been found of genetic exchange between phages from unrelated bacterial species (Krylov et al. 2013).

However, not all genes participate in horizontal transfer as is the case of the core genes which are more conserved and shared by all members of a group. These genes lie within a group of genes that "travel together through evolution", they encode for proteins with a significant biological function that interact with each other and so must remain together. These genes include those involved in DNA replication, nucleotide metabolism, and head and tail assembly (Hatfull and Hendrix 2011).

Phages also have a highly dynamic role in the environment as it has been calculated that there are about 10^{23} phage infections per second on a global scale. This suggests that the entire phage population turns over every few days, with each infection having the potential to introduce new genetic information into the host or viral particle (Hatfull and Hendrix 2011).

Phages are likely major players in the biogeochemical cycles in the biosphere. Lysis of viral hosts releases cellular carbon and nutrients, which is hypothesized to have an important impact on the cycling of organic matter in the biosphere on a global scale (Chibani-Chennoufi et al. 2004; Jover et al. 2014).

4.2.1 Phage Therapy: a new hope against antibiotic resistance

Phages can be easily manipulated and this characteristic gives them the potential to be used in biotechnology, research and medicine. Phages applications can go from diagnosis of a disease and detection of pathogenic bacterial strains using phage typing, prevention of the disease using phages as vehicles for vaccines, and disease treatment using phages as therapeutic agents against antibiotic resistant bacterial strains; they could also serve as biocontrol agents in agriculture and petroleum industry (Haq et al. 2012).

Antibiotic resistance is a critical problem in modern medicine (Frieden 2013). Every year more than two million people, in the United States alone, suffer from antibiotic resistant infections

and at least 23,000 die from these infections (Frieden 2013). Some highly antibiotic resistant infecting bacteria are *Streptococcus pneumoniae* and *Staphylococcus aureus* with 7,000 and 11,000 estimated annual deaths respectively (Frieden 2013). An escalating multiple drug resistance caused mainly by the abuse of antibiotics, and a lack of new antimicrobial approaches have reawakened the interest of phage therapy (Sulakvelidze et al. 2001).

Among the advantages of using phages as therapeutic agents is the fact that some phages may be highly strain-specific, more than antibiotics. They only infect a particular kind of bacteria, so they could be harmless to other beneficial bacteria thus reducing the possibility of opportunistic infections and it could be safe for human applications (Allen et al. 2014). Another advantage is their self-limitation, once the host is absent phages will perish, as they are obligate parasites (Haq et al. 2012).

Phage therapy may also have disadvantages: some phages may have a broader host range, phages may not be lytic under certain conditions, they may provide toxic properties to the infecting bacteria, and a strong antibody response in the patient could neutralize the phage. A solution to these problems may be the direct administration of the phage lytic enzyme, endolysin and not the virion itself. Given that phages kill specific bacterium strains this specificity could also be a disadvantage, this is why phages mixtures are mostly used to ensure the success of the therapy (Haq et al. 2012).

Phage therapy is currently being used in Russia, Poland and Georgia, but not in most western countries. There are a few exceptions, for instance in France, in Texas, and Australia, it has been used occasionally on a compassionate use basis; licensed naturopathic physicians in Washington and Oregon are permitted to use any natural product as long as it has been approved in any country in the world (Grassberger et al. 2013). In the United States the Food and Drug Administration (FDA) has not authorized phage therapy use on humans. However, in 2006 the FDA amended the food additive regulations and approved the use of a phage preparation to prevent growth of the undesirable microorganism *Listeria monocytogenes* on ready-to-eat meat and poultry products, and granted the phage preparation the status of "generally recognized as safe" (GRAS) (Verbeken et al. 2007). In 2009 the FDA approved a phage phase I clinical trial in which venous ulcers were treated with a phage cocktail targeting *S. aureus, Pseudomonas aeruginosa* and *E. coli* (Parracho et al. 2012).

4.3 Giant phages: an underrepresented group for many years

As noted before tailed phages genomes that have been sequenced varies between 3 kbp and about 500 kbp (Hatfull and Hendrix 2011). It was long thought that a genome size of 50 kbp was normal for most tailed phages and any larger genome size was an oddity (Fig 4) (Ackermann 1998). In the last ten years researchers have become aware that the standard phage isolation and propagation protocols were biased towards isolation of phage particles with a head diameter of ~60 nm which corresponds to a 50 kbp genome size (Ackermann 1998), resulting in less than 1% of the phage species in a sample being detected by plaque assays (Serwer et al. 2004).

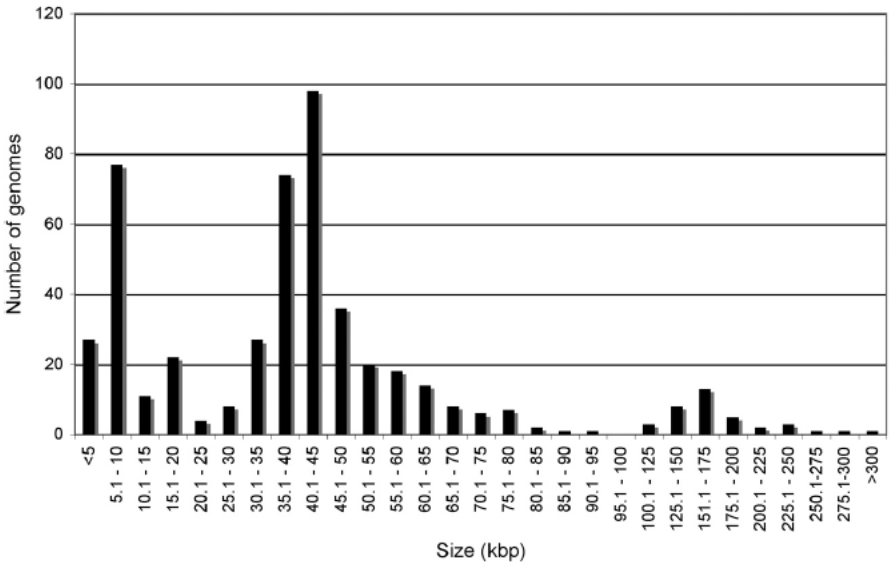

Figure 4: Sequenced bacteriophage genomes sizes at the phage database in NCBI. 500 genomes as of July 2008 (Hatfull 2008).

It has been shown that to isolate phages with larger particle sizes (required to package larger genomes) a lower percentage agar overlays (upper layer agar gels) need to be used. In plates with the standard agar overlay these larger phages are frequently not detectable at all as they form tiny plaques. The plaque size depends on the percentage of the agar gel, a reduction in the percentage will increase the size of the phage plaques allowing detection of phages with large virions (Fig 5) (Serwer et al. 2007; Kropinski et al. 2009).

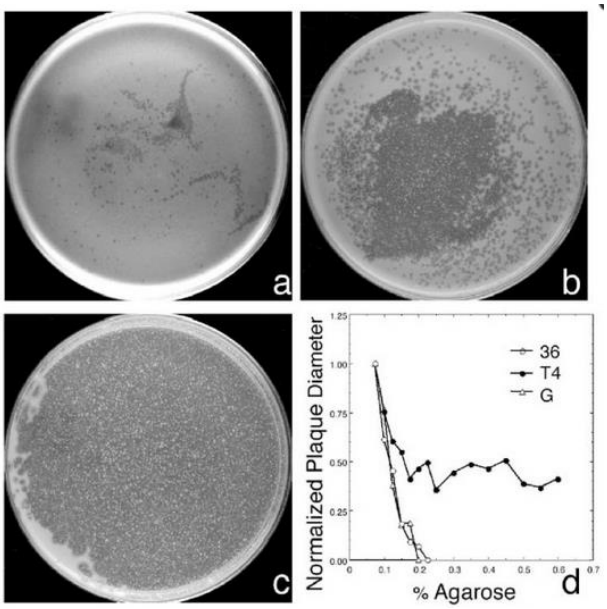

Figure 5: Concentration of agar in the overlay layer and plaque formation of phage 0305ϕ8-36. Overlay agar percentages: A) 0.4%, B) 0.2%, C) 0.15% and D) Comparison of plaque diameter at different overlay agar percentages for phages G, T4 and 0305ϕ8-36 (abbreviated by 36) (Reproduced with permission from Serwer et al. 2007).

Tailed phages which possess genomes larger than 200 kbp and correspondingly large virions are usually referred to as "Giant Phages", 49 giant phages have been sequenced to date (September, 2015). The majority of giant phages belong to the *Myoviridae* family (Table 1).

Lower costs and more advanced sequencing techniques (High-throughput next generation sequencing and modern bioinformatics tools) have triggered a revolution in phage research. In the last years a remarkable number of complete phage genomes of evolutionary or biological interest have been sequenced increasing the awareness of the total number of phages in the environment and the knowledge about phage diversity (Fig 2). This creates even more interest in sequencing new phages to compare them to known genomes. The long genomes of giant phages encode many genes, for most of them the function is unknown and they have no homologs in databases (Comeau et al. 2008; Hatfull and Hendrix 2011).

Table 1. Phages with genome lengths greater than 200 kbp. Giant ϕKZ-related phages are shown highlighted.

Year	Morphotype	Phage Name	Genome length, bp	Host	GenBank identifier	Country of Isolation
2011	myovirus	phage G	497,513	*Bacillus*	GI:347449230	USA
2012	myovirus	vB_CsaM_GAP32	358,663	*Cronobacter*	GI:414086915	Canada
2014	myovirus	121Q	348,532	*Escherichia*	GI:712914205	USA
2012	myovirus	PBECO 4	348,113	*Escherichia*	GI:441462392	Korea
2014	myovirus	K64-1	346,602	*Klebsiella*	GI:849122338	Taiwan
2012	myovirus	vB_KleM-RaK2	345,809	*Enterobacteria*	GI:375281080	Lithuania
2008	myovirus	201ϕ2-1	316,674	*Pseudomonas*	GI:189490161	USA
2010	myovirus	ϕPA3	309,208	*Pseudomonas*	GI:334737994	UK
2011	myovirus	OBP	284,757	*Pseudomonas*	GI:371671369	Belgium
2012	myovirus	Lu11	280,538	*Pseudomonas*	GI:388684650	Belgium
2001	myovirus	ϕKZ	280,334	*Pseudomonas*	GI:29134936	Belgium
2012	siphovirus	CcrColossus	279,967	*Caulobacter*	GI:414088056	USA
2013	myovirus	Ea35-70	271,084	*Erwinia*	GI:583926848	Canada
2011	myovirus	phiR1-37	262,391	*Yersinia*	GI:358356482	Finland
2013	myovirus	PaBG	258,139	*Pseudomonas*	GI:530787714	Russia
2010	myovirus	P-SSM5	251,013	*Prochlorococcus*	GI:508551825	USA
2005	myovirus	P-SSM2	252,407	*Prochlorococcus*	GI:265525079	USA
2015		ValKK3	248,088	*Vibrio*	GI:768215381	Malaysia
2010	myovirus	nt-1	247,489	*Vibrio*	GI:514050383	USA
2012	myovirus	VH7D	246,964	*Vibrio harveyi*	GI:589286276	China
2011	myovirus	phi-pp2	246,421	*Vibrio*	GI:394774562	Taiwan
2003	myovirus	KVP40	244,834	*Vibrio*	GI:91214232	USA
2012	myovirus	ϕEaH2	243,050	*Erwinia*	GI:431810571	Hungary
2011	myovirus	SPN3US	240,413	*Salmonella*	GI:349502711	South Korea
2012	myovirus	VP4B	236,053	*Vibrio*	GI:432142535	China
2003	myovirus	phage 65	235,229	*Aeromonas*	GI:326536409	USA
2003	myovirus	Aeh1	233,234	*Aeromonas*	GI:33414610	France
2010	myovirus	S-SSM6a	232,883	*Synechococcus*	GI:460109153	USA
2009	myovirus	S-SSM7	232,878	*Synechococcus*	GI:326783659	USA
2012	myovirus	CC2	231,743	*Aeromonas*	GI:423261416	China
2009	myovirus	RSL1	231,255	*Ralstonia*	GI:269838879	Japan
2013	myovirus	ACG-2014f	225,436	*Synechococcus*	GI:723041874	USA
2010	myovirus	phiAS5	225,268	*Aeromonas*	GI:310722466	South Korea
2012	myovirus	CR5	223,989	*Cronobacter*	GI:514051312	South Korea
2014	myovirus	RSL2	223,932	*Ralstonia*	GI:734703321	Japan
2012	siphovirus	CcrRogue	223,720	*Caulobacter*	GI:414088861	USA
2015	myovirus	RSF1	222,888	*Ralstonia*	GI:906846885	Japan
2010	myovirus	PX29	222,006	*Aeromonas*	GI:312262424	USA
2012	siphovirus	CcrKarma	221,828	*Caulobacter*	GI:414089228	USA
2012	myovirus	PAU	219,372	*Sphingomonas*	GI:380885820	USA

2012	siphovirus	CcrSwift	219,216	*Caulobacter*	GI:414089586	USA
2007	myovirus	0305phi8-36	218,948	*Bacillus*	GI:156563990	USA
2012	siphovirus	CcrMagneto	218,929	*Caulobacter*	GI:414088509	USA
2013	myovirus	ϕEaH1	218,339	*Erwinia*	GI:589287449	Hungary
2012	siphovirus	phiCbK	215,710	*Caulobacter*	GI:414087713	USA
2004	myovirus	EL	211,215	*Pseudomonas*	GI:82700933	Belgium
2010	myovirus	S-SKS1	208,007	*Synechococcus*	GI:472340845	North Atlantic
2015	myovirus	phiN3	206,713	*Sinorhizobium*	GI:817034960	USA
2011	myovirus	JM-2012	167,292	*Vibrio*	GI:389060148	South Korea

As previously mentioned other important factor for the increase in giant phage research has been the appropriate propagation procedures that has allowed better and more giant phage isolations, i.e. community sequencing, liquid enrichment culture and appropriate percentage of overlay agar (Serwer et al. 2007).

4.3.1 Giant ϕKZ-related phages: unique structures and genes

ϕKZ, the type virus of this group, has a circular genome of 280 kbp and infects the gram-negative pathogen *Pseudomonas aeruginosa* (Mesyanzhinov et al. 2002). *P. aeruginosa* is responsible for the most common infections in hospitals with a mortality rate of 18 to 60%, it causes pneumonia, tissue or blood infections (www.cdc.gov/hai/organisms/pseudomonas.html). Phage ϕKZ has been included in phage therapy cocktails aimed to fight pathogenic infections caused by bacteria with high antibiotic resistance (Krylov et al. 2007).

These ϕKZ-related phages are giant phages that belong to a very distant branch of the *Myoviridae* family (highlighted in Table 1) (Krylov et al. 2007). ϕKZ has a varied level of homology when compared with other phages infecting common Pseudomonad hosts as well as different Gram-negative bacteria from the genera *Cronobacter, Erwinia, Salmonella, Vibrio* and *Yersinia* (Ceyssens et al. 2014). For example extensive analysis of the ϕKZ prohead protease has demonstrated homology with proteins of phages such as *Erwinia* phage ϕEaH2, *Vibrio* phage JM-2012 and *Salmonella* phage SPN3US, our phage of research (Thomas and Black 2013).

ϕKZ-related phages are characterized for presenting a number of features highly unusual for tailed phages. They have unusual large genomes virions with a high degree of structural

complexity (more than 60 different proteins forming the virions), locating them amongst the most structurally complex prokaryotic viruses. Reassortment of structural genes in clusters along the genome has been observed due to homologous recombination as opposed to other dsDNA tailed phages which have morphopoietic genes generally grouped together (Ackermann 1998).

Their large isometric head T = 27 contains a protein cylinder known as inner body of undefined function (Fig 6). It is believed to be multi-functional and be related to DNA packaging and organization, and as it has been demonstrated by electron microscopy (EM) studies that after infection the inner body is no longer present in the head, then probably it takes part in the DNA ejection, phage development or host infection (Krylov et al. 1984; Thomas et al. 2012; Thomas and Black 2013).

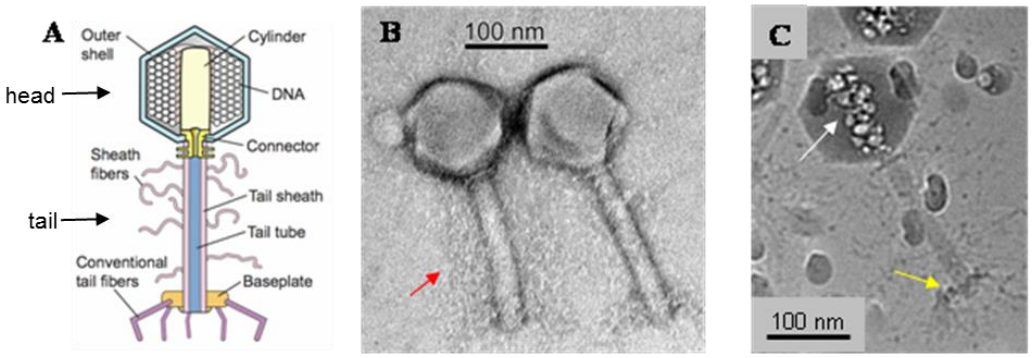

Figure 6: Morphology of a type of tailed phage with contractile tail (a ϕKZ-related phage). A) Virion structure of the ϕKZ-related phages, B) Transmission Electron Microscopy (TEM) and C) Cryo-Electron Microscopy (Cryo-EM) of ϕKZ-related phage 201ϕ2-1 (Reproduced with permission from Thomas et al. 2008) (c).

Other important and unusual characteristic of the giant ϕKZ-related phages is that they possess two sets of fragmented RNAP β and β'-like proteins which are distinct from the bacterial RNAP large subunits β and β'. The eight resulting RNAP subunits are conserved in the ϕKZ-related phages, and are present all across their genome, some of them in blocks although the genes in the subunits have lost their synteny (Fig 7). Four of these subunits are virion-associated proteins (vRNAP) and are products of middle and/or late genes. The other four are not associated with the virion (nvRNAP) and are encoded by early genes. These RNAP subunits are also found in the *Bacillus* phage PBS2, in which it is suggested that the vRNAP proteins are injected into the cell for expression, the RNAP completely assembled would be too large to pass through the tail tube.

It is known that in ϕKZ the subunits are injected and are able to infect the host without using the host's transcriptional machinery, it has been hypothesized that this also occurs in the ϕKZ-related phages. Theses phages possess a complex transcriptional program, it is interesting how the RNAP subunits become a monomeric enzyme expressed from different parts of the genome (Thomas et al. 2008; Ceyssens et al. 2014).

The members of this group are so diverged, it appears as if they are diverging twice as fast as other phages, that the RNAP subunits could not be found by simple protein-protein BLAST searches (Thomas et al. 2008). There is also a great variation in the number of genes among the ϕKZ-related phages which suggests horizontal genetic exchange, though the presence of widely dispersed core orthologs in all the phages of these different hosts including PBS2, suggests they descend from an ancient common ancestor prior to the split of the Gram-positive and Gram-negative hosts.

ϕKZ-related phages also exhibit an unusual replication having a DNA polymerase split into two polypeptides which show similarity with the 3'-5' exonuclease domain and the polymerase domain of the T4 DNA polymerase, showing a very ancient split (Cornelissen et al. 2012).

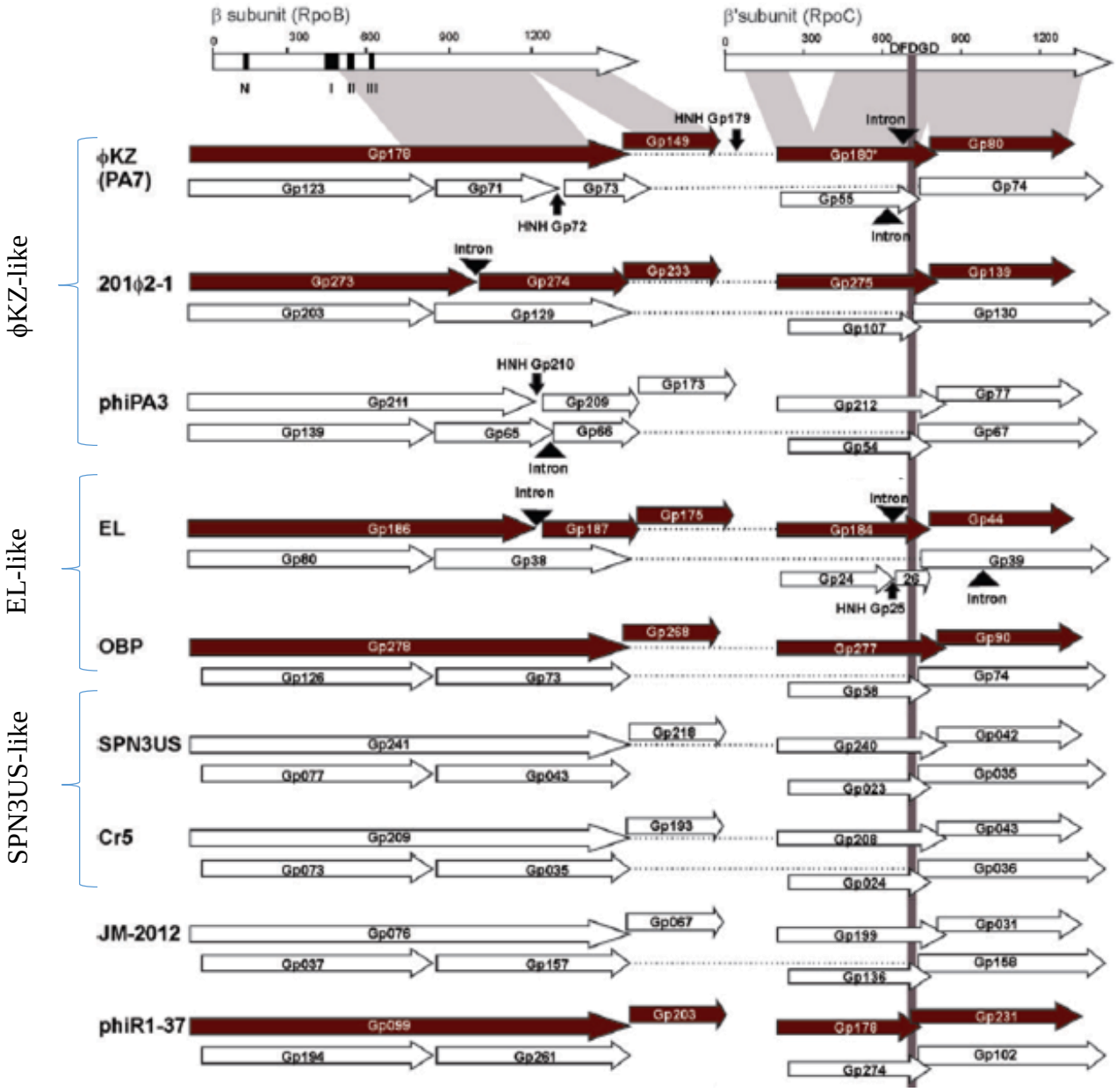

Figure 7: RNAP β and β'-like proteins of the giant ϕKZ-related phages. Proteins colored in brown are part of the virion (Adapted from Ceyssens et al. 2014).

4.4 SPN3US: our phage of research and the host *Salmonella enterica*

Salmonella enterica subspecies serovar Typhimurium is a gram-negative pathogen that causes 1.2 million cases of salmonellosis every year, which results in approximately 19,000 hospitalizations and 380 deaths (Mead et al. 1999, www.cdc.gov/salmonella/). Phages which infect this subspecies of *Salmonella* could be used as potential biocontrol agents as an alternative to antibiotics, due to the fact that a considerable proportion of the Typhimurium strains have developed resistance to several antimicrobial drugs. Phage cocktails have been proved efficient when dealing with phage-resistant *Salmonella* that uses mutation of the receptor as a route to escape phage infection (Shin et al. 2012).

In 2011 Lee et al. sequenced the complete genome of phage SPN3US which was isolated from chicken feces using *Salmonella enterica* serovar Typhimurium LT2 as a host. The genome of phage SPN3US has 240,413 bp of length with 264 ORFs clustered in both strands; however, due to insufficient database information about the *Salmonella* phage genome functional genes, 79.2% of the ORFs were annotated as hypothetical (Lee et al. 2011). The genes with an assigned function include those related to DNA packaging, phage structure, tail structure for host interaction, replication/transcription and host lysis (Lee et al. 2011).

The quantification in the relationship between phages SPN3US and ϕKZ was examined to establish that SPN3US was a good model phage to employ for the ϕKZ-related phages. Based on Lavigne et al. (2009) protein similarity approach to classify Myoviruses which employs a cutoff value of 40% homologous proteins to delineate genera and of 20-30% to delineate subfamilies it was determined that SPN3US belonged to the ϕKZ-related subfamily as 32.6% of its proteins have similarity to ϕKZ proteins. It was also observed that 81.4% and 59.1% of SPN3US proteins had similarity to proteins in *Erwinia* phage ϕEaH2 and *Cronobacter* phage CR5 respectively which tentatively classifies the three phages into a new genus. SPN3US could be considered the type phage for this new genus, the SPN3US-like phages (Thomas unpublished). The status of these newly identified relationships is expected to become clearer as new members of these phages are sequenced.

Among the similar characteristics that SPN3US shares with the ϕKZ-related phages it has been determined the structural similarity with the major capsid protein, the tail sheath protein, the tail tube protein, the presence of the RNAP subunits (Sycheva et al. 2012; Ceyssens et al. 2014;

Sokolova et al. 2014) and it is probable to find an inner body structure given the existence of proteins homologous to the ϕKZ inner body proteins (gp89 and gp93) (Thomas and Black 2013).

To the best of our knowledge, no studies have been performed with SPN3US to determine if it could be used for phage therapy.

4.5 The importance of studying phage SPN3US

Since the publication of *Salmonella* phage SPN3US complete genome in 2011 there have been no further publications, thus more research needs to be done to determine the functions of all its genes with unassigned functions.

It has been shown in several well studied model phages, such as T4, that essential genes are more conserved than non-essential genes as their name implies are necessary for the survival of the species (Comeau et al. 2007). Using amber mutant phages, which have a premature amber (TAG) stop codon disrupting a gene, we hypothesize the essential genes for SPN3US phage can be identified and their function(s) determined. With this approach we should be able to determine some essential structural genes encoding for proteins that are part of the phage virion.

The goal of the present research will be to sequence mutant phages isolated for SPN3US that from plating characteristics have an amber mutant phage phenotype as they only grow on suppressor strains of *Salmonella* (which possess a specialized tRNA able to insert an amino acid at the amber codon allowing the production of a full-length protein) and not on non-suppressor strains. We hypothesize that the sequence will show amber mutations truncating the genes of the SPN3US, and if there is one amber mutation truncating one gene in one mutant we can infer that that gene has an essential function in the phage life style.

Our hypothesis is that we can use SPN3US as a model to study all ϕKZ-related phages who have homologous genes. This is critical as the core set of genes for these ϕKZ-related phages is completely unknown. Previously such genetic studies could not be performed on the ϕKZ-related phages as there are no suppressor hosts in *Pseudomonas*, but there are suppressor hosts for *Salmonella*. There is not much information about these complex bacterial viruses, so by studying SPN3US we can use it as the first genetic model for giant phages.

We hypothesize that some of the genes to be essential in SPN3US will be part of the core genes in the ϕKZ-related phage group. However, we also hypothesize that some of the essential genes will not be well conserved in other ϕKZ-related phages as conceivably it will have essential genes that have a role specific to its own host, e.g. tail fiber genes encoding for proteins responsible for binding to the host receptor. As this is the beginning of the creation of a novel model system for studying the ϕKZ-related phages this research will also be important to help refine the methodology and analyses to determine which are the most appropriate, economical and time efficient processes. We hope to be able to blend classical genetic studies with next generation sequencing approaches in a manner that has not been used for phages previously. In the longer term, if determined to be successful, this process can be applied to other phages to identify the functions of their essential genes. This study will have broader implications in researchers also studying ϕKZ-related and other phages including researchers working on phage therapy.

5. MATERIALS AND METHODS

5.1 Isolation and characterization of amber mutants of SPN3US

5.1.1 Mutagenesis of SPN3US

The hydroxylamine random mutagenesis method described for T4 by Tessman (1968) was followed to obtain amber mutant candidates of SPN3US. Hydroxylamine adds a hydroxyl group to cytosine producing hydroxylaminocytosine, which tends to pair with adenine and results in GC to AT transitions. When such a mutation occurs in a tryptophan codon (TGG) or a glutamine codon (CAG) the result is an amber nonsense mutation (TAG codon) leading to an early termination of the protein translation (Fresse et al. 1961).

The mutagenized and control samples of SPN3US contained 500 uL of 0.1 M sodium phosphate buffer (pH 6), 2 uL of 0.5 M EDTA and 100 uL of SPN3US phage stock (high titer wild type). 400 uL of 1M hydroxylamine solution was added to the mutagenized sample, while the control sample received 400 uL of SM buffer. The samples were incubated at 37°C for 24 hours. The solutions were diluted 1:100 into LB media containing a final concentration of 1 mM EDTA and stored at 4°C (Tessman 1968).

5.1.2 Plaque assay

A plaque assay was conducted to determine the titer (number of viable phage virions) of the mutagenized and control samples. A serial ten-fold dilution up to 10^{-6} was prepared for each sample using SM buffer as the diluent. Overlays were prepared with 4 mL of soft agar (0.34% agar) in LB+N and 100 uL of non-suppressor (sup^-) strains (Table 2) in exponential phase, the mixture was laid over a hard LB agar base and spread evenly. The dilutions were spotted (5 uL) on the solidified top agars, let to dry and the plates were incubated at 30°C overnight. Plaques were counted for each dilution and the titer was calculated (plaque forming units (PFU/mL) = number of plaques / volume of phage solution x dilution factor). Plates were stored at 4°C.

Table 2. *Salmonella* typhimurium LT2 strains used in this study

Non-suppressor (sup$^-$)	**Suppressor (sup$^+$)**
UB0015	UB0017 (supE)[a]
TT9070	TT6675 (supD)[b]

[a] sup$^+$ strain with specialized tRNA inserts glutamine codon at amber codon
[b] sup$^+$ strain with specialized tRNA inserts serine codon at amber codon

5.1.3 Isolation of amber mutant candidates

The suppressor (sup$^+$) hosts (Table 2) were infected with an appropriate amount of the mutagenized sample to obtain 100 to 200 well separated plaques. Prior to plating the mutagenized phage was incubated with exponential stage host at 37°C for 15 minutes to allow for absorption. The host-phage mixture was added to the soft agar and poured onto the agar plate, let to solidify and the plates were incubated at 30°C overnight (Thomas 2015).

The plaques obtained were then tested on the sup$^-$ strains and the corresponding sup$^+$ strains to detect which plaques had an amber phenotype (infecting only sup$^+$ strains and not sup$^-$ strains in contrast with the wild-type phage which can grow on both strains). Isolated plaques were gently poked with sterile toothpicks and stabbed into a plate with a sup$^-$ overlay and then into a plate with a sup$^+$ overlay in an analogous location. This procedure was repeated for as many individual plaques that could be seen on the plate. The plates were incubated at 30°C overnight. SPN3US amber mutant candidates that grew on the sup$^+$ strain but did not grew on the sup$^-$ strain were re-screened to confirm that they retained the amber phenotype (Thomas 2015).

A crude stock of each of the SPN3US amber mutant candidates was created. Plaques that only grew on the sup$^+$ host were plugged with a sterile Pasteur pipette and placed into an Eppendorf tube with SM Buffer and 1:50 volume of chloroform. After 30 minutes of incubation at room temperature the mixture was centrifuged at 4,000 rpm for 5 minutes. The supernatant was serially diluted up to 10^{-4} and spotted on sup$^+$ and sup$^-$ plates to determine the titer and to ensure that the candidate was a true positive for an amber mutant phenotype (Thomas 2015).

To obtain a high titer stock, plates with single plaques were prepared with the 10^{-4} dilution from the crude stock. A single plaque was stabbed several times into the hard agar plate, the overlay was prepared with the sup$^+$ host (300 to 500 uL) and added to the stabbed bottom plate making

sure it was evenly spread. The plates were incubated at 30°C overnight. The overlay for each mutant candidate was harvested using a clean glass slide and 3 mL of SM buffer with lysozyme (0.2 mg/mL) was added and the mixture was left at 4°C overnight. The mixture was differentially centrifuged at 6,000 rpm for 10 minutes at 4°C (to spin down agar, host cells and debris) and the supernatant was then centrifuged at 18,000 rpm for 30 minutes at 4°C. The supernatant was removed, SM buffer was added to the phage pellet and it was left to resuspend at 4°C overnight. The high titer phage stock was transferred to a 1.5 mL Eppendorf and stored at 4°C. Dilutions of the phage stock were spotted on sup^+ plates to determine the titer and on sup^- plates to calculate the reversion rate (rr = titer sup^- /titer sup^+) (Thomas 2015).

Cross-plating tests were performed to determine if the amber mutants were genetically distinct from each other. If the amber mutants are different they will be able to rescue each other via complementation (at the protein level) or via recombination (at the DNA level). Overlays were prepared with soft agar, ~10^7 particles of each mutant and the sup^- host, also control overlays were prepared only with sup^- and sup^+ hosts. A dilution series (1:100) of each amber mutant were spotted on the above overlays. The plates were incubated at 30°C overnight (Thomas 2015).

5.2 Sequencing of SPN3US amber mutant phage genomic DNA using next generation sequencing

SPN3US amber mutant phage DNAs were extracted and purified from single plaque-derived high titer stock solutions (~10^{11} pfu/mL). Some amber mutant phage stocks were from a collection which had previously been isolated in the serine suppressor host, and undergone growth characterization studies (Thomas and Black unpublished) and five amber mutant phages have been newly isolated in this project as described in 3.1. Some phage stocks were combined in the ratios provided in Table 3 to determine if sequencing two ambers together was feasible to identify the mutation sites.

Genomic phage DNA was extracted using a Phage DNA Isolation Kit (Norgen Biotek Corp., Thorold, Canada) using the manufacturer's recommended protocol with slight modifications. Typically, 50 to 100 uL of the concentrated phage suspension were used and the DNase inactivation step was omitted. The isolated DNAs were run on a 1% agarose gel to quantify the

DNA yield against the GeneRuler 1 kbp DNA ladder (Fermentas). Illumina sequencing was performed at the University of Rochester Genomics Research Center. Libraries were generated with a Nextera XT kit and paired-end reads were obtained with an Illumina HiSeq2500 sequencer for the first group of samples and with an Illumina MiSeq for the second group of samples. The sequenced reads were cleaned using a rigorous pre-processing workflow (Trimmomatic-0.32, www.usadellab.org/cms/?page=trimmomatic). The adapters were removed, a sliding window of 4 base length was used to scan the reads and cut where the average quality per base falls below 20, the leading and trailing low quality or N bases with quality score below 13 were removed, and the reads with less than 25 bases were dropped. A quality control analysis was performed with FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc/). This was done by University of Rochester Genomics Research Center.

Table 3. Amber mutant phage sequencing samples

Sequencing System	Sample	Ambers	Ratio
HiSeq		18	
		19	
	F	21/28	1:6.25
	G	23/29	1:6.25
	H	24/30	1:6.25
	I	25/31	1:6.25
	J	1	
	K	6	
	L	13	
	N	20	
	O	26	
MiSeq		107	
		108	
		109	
		110	
		111	

5.3 Determination and confirmation of amber mutation sites in SPN3US mutant phage genomes

The Illumina fastq sequence reads were aligned to the SPN3US reference sequence (NCBI accession number GI:349502711) using the templated alignment in SeqMan NGen sequence assembly software (DNASTAR, Madison, WI). The coverage was evaluated and the SNPs were identified and analyzed using SeqMan Pro (DNASTAR, Madison, WI). For the first set of samples sequenced using the HiSeq system, once the location of the amber mutation sites was determined the regions flanking the mutation sites underwent amplification using PCR. These PCR products were purified and used as the substrate for Sanger sequencing to confirm the position of the mutation.

5.3.1 PCR of SPN3US genomic regions containing putative amber mutations

PCR was performed on the genes of interest (ones with an identified SNP) from both the wild type SPN3US phage (as a positive control) and from the relevant amber mutant(s), using primers synthesized by Life Technologies. The fasta file of the genes of interest were downloaded from the GenBank, if the gene of interest was coded in the complementary strand then the sequence was reversed and complemented. The codon with the mutation site was located using the coordinates from the SeqMan Pro SNP report. For the primer design an online Oligonucleotides Properties Calculator (www.basic.northwestern.edu/biotools/oligocalc.html) was used making sure the GC content was between 40 and 60%, the melting temperatures for both primers were similar and the length of the primers ranged from 25 to 35 nt. For some PCR products the primer design included restriction endonuclease sites so that those PCR products could potentially not only be used for confirmatory sequencing (section 3.3.2) but also for future cloning purposes, such as expression of the protein of interest or recombination analysis. Restriction sites were determined using the restriction mapping tool Webcutter 2.0 (http://rna.lundberg.gu.se/cutter2/).

Typically, a PCR mix (50 uL) would contain the following components: 25 uL of 2X reaction mix (One*Taq*® Hot Start Quick-Load® 2X Master Mix with GC Buffer from New England BioLabs), 2 uL of template (~10^6 pfu), 0.5 uL of each primer (final concentration of 1 μM), and 22 uL of DEPC treated water. The amplifications were performed in a Veriti 96-well Thermal Cycler, using

temperature grading PCR to identify the best annealing temperature to optimize the reaction and obtain the best product yield. The amplification programs typically consisted of: an initial denaturation step at 94°C for 2 min, 30 cycles of 3 steps (denaturation step at 94°C for 15 sec, annealing step at temperatures ranging from 56 to 60°C for 30 sec, and an extension step at 68°C for one min per kbp of predicted product length plus an additional minute), a final extension at 68°C for 7 minutes and an indefinite period at 4°C.

PCR products underwent electrophoresis on a 1% agarose gel containing ethidium bromide (final concentration of 0.5 ug/mL) in TAE buffer at 90-100V for 30-40 minutes, in order to confirm the amplification was successful, the product was the correct size, no other products were amplified and also to estimate the DNA concentration. The GeneRuler 1 kbp DNA ladder (Fermentas) was used as a molecular weight marker. PCR products were purified using the QIAquick PCR purification kit (QIAGEN) using the manufacturer's recommended protocol and the DNA concentration was measured with the Thermo Scientific NanoDrop 2000 Spectrophotometer.

5.3.2 Confirmatory sequencing using Sanger sequencing technology

An internal sequencing primer was designed ~100 residues upstream of the mutation sites and synthesized by Life Technologies. Samples for Sanger sequencing were prepared with 5 uL (25 pmol) of the sequencing primer and 10 uL (~20 to ~40 ng depending on the template length) of the purified PCR product. Sanger Sequencing was performed at GENEWIZ.Inc (South Plainfield, NJ). Chromas Lite was used to visualize the sequence chromatograms and to obtain the sequence in FASTA format. In order to confirm the mutation site for each amber mutant phage the sequenced regions were aligned against the reference SPN3US phage genome using the BLASTn search tool at NCBI with the default settings.

5.4 Homology relationships of identified essential genes of SPN3US using Bioinformatics tools

The workflow detailing the methods used to analyze the homology relationships is illustrated in figure 8.

Applications from the standalone BLAST+ suite were used via command line in the Linux server to perform the homology analyses (Camacho et al. 2009).

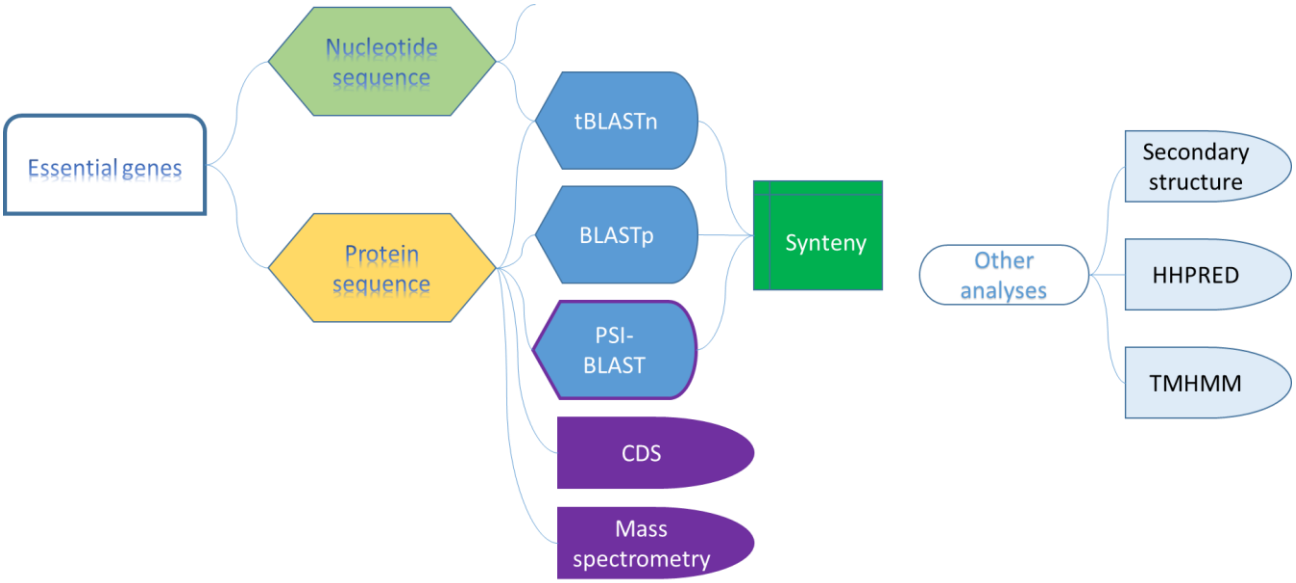

Figure 8: Flowchart outlining the strategy and programs employed to find true hits for the SPN3US identified essential genes

5.4.1 Homology of the nucleotide sequences of the SPN3US identified essential genes against ϕKZ-related genomes

Complete genomes from the ϕKZ-related phages (highlighted in Table 1) were downloaded (May 2015) from the GenBank database (www.ncbi.nlm.nih.gov/GneBank/index.html) except *Ralstonia* phage RSF1 as it was added in the GenBank database in July 2015. The thirteen genomes were concatenated into one FASTA file using the command 'awk'. A BLAST database was created using the command 'makeblastdb' specifying the sequence type (-dbtype nucl).

The eleven newly identified SPN3US essential genes and three additional control genes with known function (gp260 terminase, gp256 sheath and gp75 major capsid) were each used as queries for searches against the newly created BLAST database. The command employed was 'blastn', specifying the query sequence (-query), the subject database (-db), the expected value threshold (-evalue) and the output format (-outfmt '6 std stitle') which results in the default 12 column tabular output ('qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore') with the name of the subject organism.

An e-value cutoff of 1e-3 was used as a stringent delimiter, and broader e-value cutoffs of 1 and 10 were also employed to determine if additional more diverged relevant genes could be found.

A program was written to extract the data from the output txt file and create a comma separated values (csv) file. The results were then analyzed on excel to determine which were true homologs.

5.4.2 Homology of the identified essential protein sequences of the SPN3US against the ϕKZ-related protein database

Protein BLAST databases were constructed from the ϕKZ-related phage proteins (downloaded on May 2015 from NCBI) using the command 'makeblastdb' and the sequence type (-dbtype prot).

The gene products corresponding to the eleven essential genes and the three control genes from phage SPN3US were subjected to BLASTp searches against the ϕKZ-related protein BLAST databases using the command 'blastp' and the same parameters specified above for BLASTn, except for the expected value. The e-value cutoff of 1e-4 was used as a stringent delimiter, and broader e-value cutoffs of 1 and 10 were set to determine if additional homologous proteins could be found.

The same strategy was used to run a comparison between the protein sequences against the translated genome sequences (from all frames) using the command 'tblastn'.

The SPN3US proteins were also compared with the proteins in the non-redundant GenBank databases (downloaded April 2015) by using a PSIBLAST search with the command 'psiblast' with four iterations (-num iterations).

A program was written to extract and merge the data from the resulting txt files and a csv file was created.

5.4.3 Analysis of gene synteny between ϕKZ-related phages

Three parameters were considered to determine the significance of the obtained hits: e-value, percentage of identity, query coverage and position in the genome. The best hits with lower e-value, highest percentage of identity and query coverage were considered proteins with reliable sequence similarity.

The presence or absence of syntenous regions were analyzed for phage SPN3US and the related phages for each identified essential gene. Given that more than one hit was present for the same essential gene or no hits were found with the stringent e-value cutoff, the position of the hits resulting from the alignments with e-values 1 and 10 was determined to verify if it was located inside the synteny region and with higher confidence establish that it is a reliable hit.

Once the reliable hits were determined, if the genes in the same loci in the related phages had an assigned function, a prediction of the function for the identified essential genes was established.

Genome maps were constructed using the GenomeDiagram module of Biopython (Cock et al. 2009).

5.4.4 Conserved domain search of the identified essential proteins

A conserved domain search CDS (Marchler-Bauer et al. 2005) was performed for each identified essential protein and the control proteins on the NCBI online platform using the default settings with an e-value threshold of 0.01.

5.4.5 Secondary structure predictions of the identified essential proteins

In the absence of a significant sequence similarity the secondary structure was predicted using the online tool PSIPRED (Jones 1999) and the structural similarity with the query protein was determined.

5.4.6 Analysis of mass spectral data from the wild-type phage

Essential genes that encode proteins that are part of the phage virion were determined via analyses of mass spectral data obtained from purified and concentrated wild-type phage conducted using the CsCl ultracentrifugation method as described by Thomas et al. (2008). The purified phage was boiled in Laemmli gel loading buffer and sent to the University of Texas Health Science Center Mass Spectrometry Core Facility where it underwent GeLCMS technique.

To determine the presence of capsid proteins that are subject to cleavage by the phage protease the cleavage pattern of related phages was considered and the peptide coverage for each SPN3US virion protein was examined manually to identify any semi-tryptic peptides cleaved at a similar motif (Thomas et al. 2010).

6. RESULTS

6.1 Isolation of amber mutant candidates of SPN3US

Six SPN3US amber mutant candidates were obtained following the hydroxylamine mutagenesis treatment. A 125-fold decrease in the number of PFU/mL was observed in the mutagenized sample ($8x10^6$ PFU/mL) when compared with the control sample ($1x10^9$ PFU/mL) (Fig 9a).

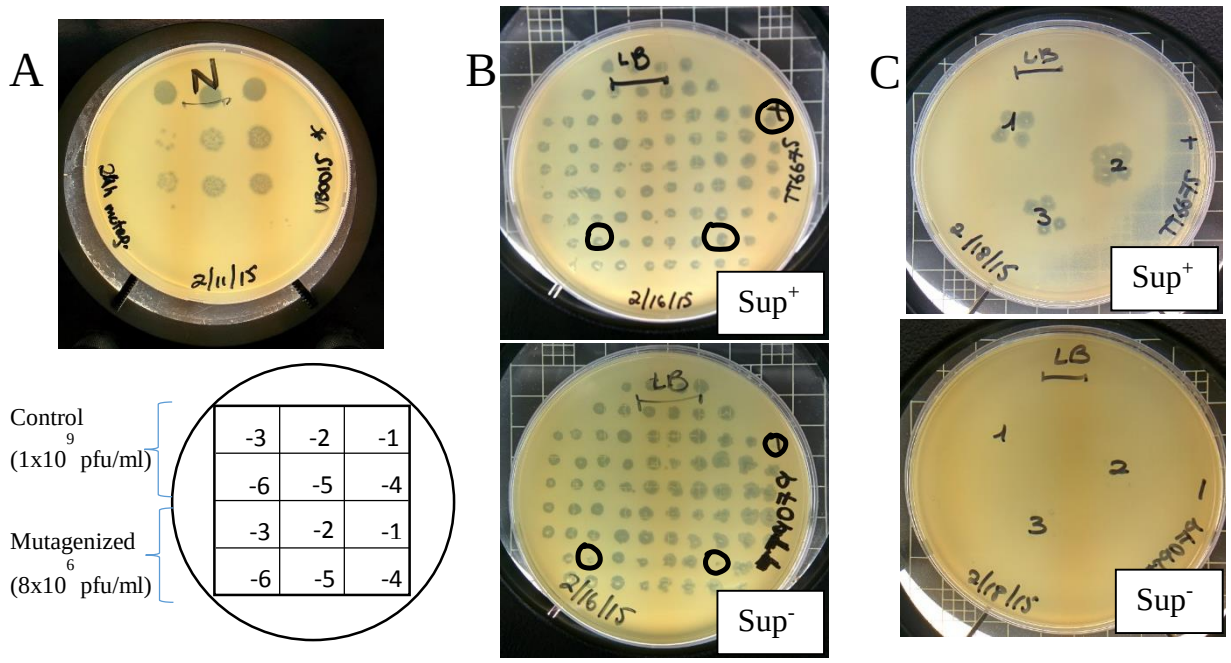

Figure 9: SPN3US amber mutant candidates' isolation and screening. A) Plaque assay using ten-fold dilutions and titer comparison between control and mutagenized samples, B) Screening of isolated plaques from the mutagenized mix in sup^+ and sup^- hosts, C) Re-screening of possible amber mutant candidates (showing zone of clearing in the sup^+ host and not in the sup^- host).

Several plates were prepared with the host-mutagenized phage mixture to obtain isolated plaques. Approximately 300 isolated plaques were screened to detect which ones had an amber phenotype, from these nine showed a zone of clearing in the sup^+ host and not in the sup^- host and were considered possible amber mutant candidates (Fig 9b). These nine plaques were re-screened and only six plaques produced a zone of clearing different enough on both hosts to consider that they held true the amber phenotype (Fig 9c).

Six crude stocks were prepared (D1, D2, D3, B1, B3 and B4) and the plaque assay showed that the amber mutant candidate B3 was mixed with wild type phage and required an additional screening. The titers for the six concentrated stocks were determined (between $1x10^{12}$ and $8x10^{12}$ pfu/ml) and the reversion rates were calculated (between $2x10^{-9}$ and $1x10^{-6}$).

The cross-plating test showed that amber mutants D1, D2, D3 and B4 were able to rescue each other demonstrating that they were genetically different. As can be seen in Figure 10 amber mutants B1 and B3 were unable to rescue one another thus it was considered that they likely contained the same amber mutation. When a pair of amber mutants lead to a productive infection of the sup- host it is said to be complementing, as contrast to a pair of noncomplementing amber mutants which do not lead to active progeny particles (Epstein et al. 2012).

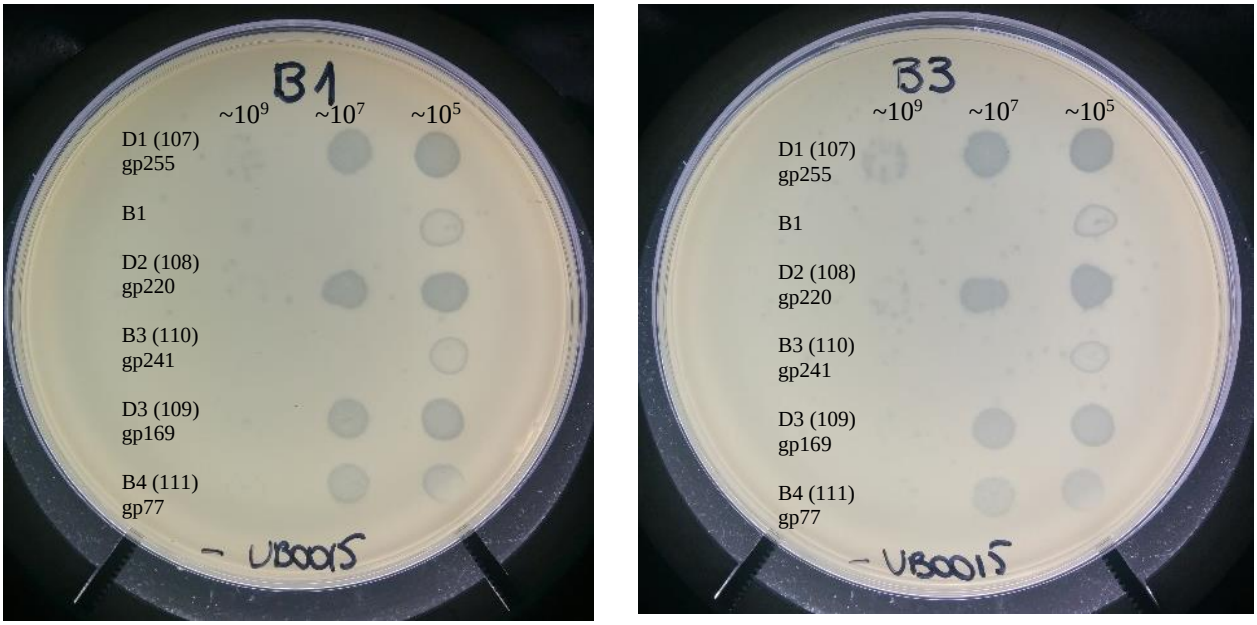

Figure 10: Cross-plating test of the amber mutant candidates of SPN3US. B1 and B3 amber mutant candidates were plated on the sup^- host (UB0015), the mutant candidates (D1, B1, D2, B3, D3 and B4) were then spotted on the overlay. B1 and B3 were unable to infect the sup^- host which contained either amber mutant in the overlay. The amber mutant number is written between parentheses and the protein in whose gene the mutation was found is written below.

The five amber mutants, D1, D2 and D3 isolated in the serine suppressor host, and B3 and B4 isolated in the glutamine suppressor host (Table 2) were added to the amber mutant collection of the Phage Laboratory and given the numbers 107 to 111 respectively.

6.2 Genome sequence results of the amber mutant candidates of SPN3US

A mix of 50 amber mutant candidates, each amber candidate added in equal amounts, had been previously sequenced on an Illumina HiSeq by LC Sciences (Houston, TX) with the objective of verifying if nonsense amber mutations were present. Approximately 77 million reads were obtained per sample. The analysis and identification of the amber mutations was performed by Accura Science. A total of 31 amber mutations were identified, 24 with CAG codons and 7 with TGG codons (Table 4).

Table 4. Nonsense amber mutation sites found for a mix of 50 amber mutant phages previously sequenced, 31 mutations were identified. Twelve of the previously identified mutation sites were verified in the present study (shown in blue).

Genome position	Refer. Base	Alter. Base	Refer. Number	Alter. Number	Alter. freq. %
215999	C	T	28278	3645	12.9
23427	C	T	17315	1776	10.3
195232	C	T	33887	2181	6.4
169242	C	T	22789	1374	6.0
186856	G	A	22375	1240	5.5
37455	C	T	31428	1439	4.6
192318	G	A	25506	995	3.9
33261	C	T	35090	1295	3.7
30811	C	T	37702	1370	3.6
63849	C	T	40330	1415	3.5
46330	C	T	32816	1092	3.3
201073	C	T	36575	1189	3.3
66658	C	T	28668	916	3.2
153240	C	T	25241	785	3.1
23982	C	T	34630	1054	3.0
230169	G	A	17353	499	2.9

Genome position	Refer. Base	Alter. Base	Refer. Number	Alter. Number	Alter. Freq. %
13997	C	T	24577	597	2.4
203975	G	A	21902	513	2.3
159924	C	T	18225	420	2.3
20159	C	T	28444	612	2.2
190698	C	T	34501	731	2.1
146047	C	T	31761	633	2.0
159099	C	T	48284	898	1.9
63957	C	T	41325	638	1.5
170989	G	A	30758	416	1.4
179577	C	T	26162	322	1.2
179287	G	A	29504	352	1.2
77022	C	T	33542	239	0.7
81829	G	A	37167	219	0.6
229576	C	T	33223	183	0.6
170544	C	T	30232	126	0.4

The next step was to determine to which amber mutant candidate each mutation(s) corresponded to. An additional goal was to determine if preparing a mix of two amber mutant candidates for sequencing would give good results and be more economical. From this group of amber mutant candidates sent in this study (Table 3, HiSeq section) the total number of sequences processed varied from 33 to 44 million per sample with a sequence length of 125 base pairs and an overall GC percentage of 48% of all bases in all sequences which corresponds with the GC content described in the SPN3US genome announcement (Lee et al. 2011). Overall the quality of the high throughput sequence data was good as determined by the quality scores per base sequenced, which remained high all the way across the runs. A high sequence duplication level (between 50 and 500 repeats) was detected for close to 50% of the sequences.

As it will be described in more detail in the next section, the sequencing was successful for two of the combined samples (Table 5.1); however, for the other two samples not all the nonsense mutations sites could be identified. In order to have a good coverage and more reliable data the sequencing core facility suggested we should try a more economical system using MiSeq, so a final group of individual amber mutant candidates (Table 3, MiSeq section) was sent to be sequenced.

6.2.1 Sequence assembly and SNP analysis

The sequence assembly of the reads obtained through the HiSeq system took around three hours to be completed while the assembly of the reads obtained through the MiSeq system took around two minutes. The assemblies were performed in an Apple desktop with a Window operating system, an Intel Xeon Processor, 2.80 GHz of speed and 22 GB of RAM. Given that the data was aligned using a template sequence there was only one contig (lengths varying from 242900 to 243724). Following the assembly in SeqMan NGen the project was viewed in SeqMan Pro and the coverage and SNPs were analyzed. For all the samples the great majority of the contig exceeded the maximum depth of coverage (default parameter 100), this can be easily visualized in Figure 11 where the areas exceeding the coverage parameter are shown in red.

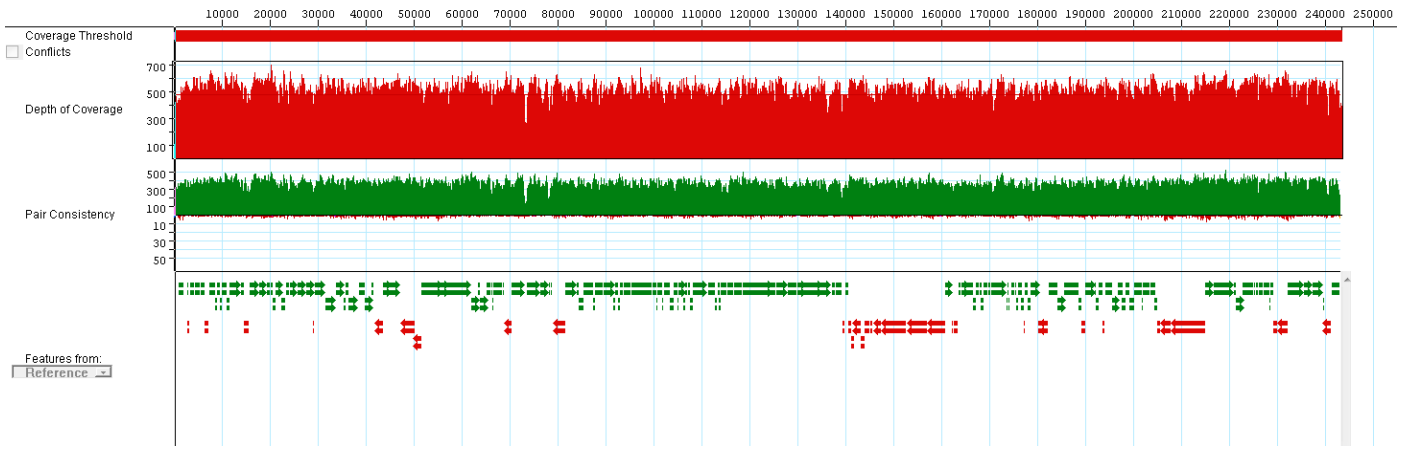

Figure 11: Strategy view in SeqMan Pro showing the depth of coverage for the contig of the amber mutant 107. The majority of the contig was colored in red as it was well above the default parameter (100).

The SNP summary report for the contigs of each sample showed that various putative SNPs were identified (Appendix A). Special focus was given to the nonsense mutations given that the amber mutations are nonsense. The depth of coverage for the SNPs presenting the nonsense mutation with the MiSeq system were between 400 and 900 fold. Table 5.1 and 5.2 list the nonsense mutations found in each sample, the position of the mutation and the corresponding gene product. From the combined samples submitted for sequencing (Table 3) only samples F and H showed two nonsense mutations with the SNP percentage having a direct correlation with the ratios in which the samples were combined. In sample I only one nonsense mutation was detected with the SNP occurring at nearly 100%, though two nonsense mutations were expected, and in mix G no nonsense mutations were detected though two were expected. Two nonsense mutations were identified in amber mutant 26, making it a double mutant, and a single nonsense mutation was found in the rest of the samples. Amber mutants 18 and 19 had mutations in different positions in the same gene (gp203), and the same could be seen for amber mutant 109 and sample I with different amber mutations seen in gp241. Amber mutation phages 6 and 20 had a mutation in the same position in the same gene (gp25).

From the fourteen identified essential genes, only three had previously been assigned functions (GenBank accession number JN641803.1) (Table 5.1 and 5.2).

Table 5.1 Amber nonsense mutations identified in the amber mutant candidates sequenced by HiSeq system

Sample[a]	Amber mutants	GP	Length (AA)	Predicted function	SNP Pos.	DNA change[b]	Codon[c]	AA change[d]	SNP %[e]
	18	203	459	putative virion protein	179577	67:G>A	TGG:2	W:126	
	19	203	459	putative virion protein	179287	56:C>T	CAG:1	Q:223	
F	21 28	214 238	251 740	hypothetical hypothetical	186856 203975	682:C>T 1186:C>T	CAG:1 CAG:1	Q:228 Q:396	**85.70%** **13.80%**
G	23 29								
H	24 30	171 219	420 242	hypothetical hypothetical	159099 190698	361:C>T 271:C>T	CAG:1 CAG:1	Q:121 Q:91	**82.30%** **15.70%**
I	25 31	241	1401	putative DNA- directed RNApol β	215999	2011:C>T	CAG:1	Q:671	99.80%
J	1	47	573	hypothetical	46330	1432:C>T	CAG:1	Q:478	99.80%
K	6	25	128	hypothetical	20159	184:C>T	CAG:1	Q:62	99.50%
L	13	70	286	hypothetical	66658	181:C>T	CAG:1	Q:61	99.70%
N	20	25	128	hypothetical	20159	184:C>T	CAG:1	Q:62	99.70%
O	26	19	93	hypothetical	13997	70:C>T	CAG:1	Q:24	99.50%
		168	1727	hypothetical	146047	4673:G>A	TGG:2	W:1558	99.80%

[a] Samples F, G, H and I each contained two amber mutants combined in a 1:6.25 ratio (underlined amber was included in highest amount)

[b] Denotes in which nucleotide the change occurred, the reference base and the new base.

[c] Codon in the reference sequence and the position where the change occurred

[d] Amino acid in the reference sequence and position

[e] Percentage of the most predominant non-reference base in the SNP position

Table 5.2 Amber nonsense mutations identified in the amber mutant candidates sequenced by MiSeq system

Amber mutant	GP	Length (AA)	Predicted function	SNP Pos.	DNA change	Codon	AA change	SNP %
107	255	291	Putative virion protein	226689	448:C>T	CAG:1	Q:150	99.80%
108	220	156	hypothetical	191499	76:C>T	CAG:1	Q:26	100%
109	241	1401	putative DNA- directed RNApol β	214121	133:C>T	CAG:1	Q:45	100%
110	169	1376	hypothetical	152931	2015:G>A	TGG:2	W:672	99.60%
111	77	592	hypothetical	74357	604:C>T	CAG:1	Q:202	99.80%

6.2.2 Confirmation of mutation sites with Sanger sequencing

Since the identification of amber mutations was not clear in all of the samples that underwent HiSeq sequencing, we decided to confirm the identified amber mutation sites in those samples for which such a mutation had been detected (Table 5.1) using Sanger sequencing. The designed primers are listed in Table 6. PCR products were amplified from the actual mutant phages samples that had previously undergone the DNA extraction for the HiSeq Illumina sequencing (see Fig 12 for an example of gp171 and 219). For each gene amplified a wild-type control PCR was also included. PCR products were examined for each sample to ensure the observed molecular weight was consistent with expected and that there were no non-specific products. There were no major differences in yield between the PCR products resulting from the gradient temperatures, so an annealing temperature of 56ºC was employed for amplifying the products that underwent Sanger sequencing.

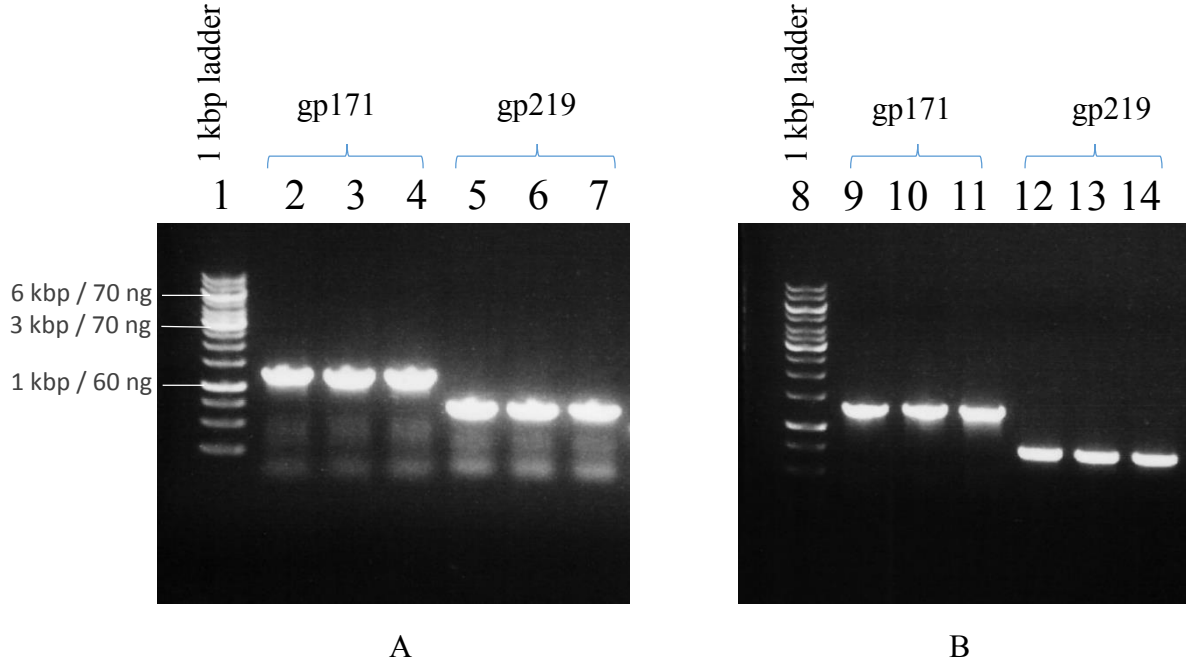

Figure 12: PCR products from amber mutant candidates 24 and 30. A) Not purified PCR products: lane 1, GeneRuler 1 kbp DNA ladder (500 ng); lanes 2-4, PCR products from amber 24, wild-type and amber 30 amplified with primers for gp171 (5 uL); lanes 5-7, PCR products from amber 24, wild-type and amber 30 amplified with primers for gp219 (5 uL). B) Purified PCR products: lane 8, GeneRuler 1 kbp DNA ladder (200 ng); lanes 9-11, purified PCR products from lanes 2-4 (2 uL); lanes 12-14, purified PCR products from lanes 5-7 (2 uL).

Table 6. Primers designed to amplify regions flanking the mutation sites and for Sanger sequencing

Gene	Primer name[a]	Sequence	GC content	Melting temperature (°C)
gp19	pH_S19F	GCG CCA TGG ATC TGA ACG TGA TGG AG	58%	62.7
	pH_S19R	GCG TCT AGA CGT TTC GCC TGC CTG AT	58%	62.7
	S19seq	GCG GAG GAG TAA GAA TAA GTC ACG ATC G	50%	61.4
gp25	pH_S25F	GCG CCA TGG TTG CGG TAG GAA CCT TAT	56%	62.8
	pH_S25R	GCG TCT AGA TTA CAT ATC CTC AAT CGC ACT C	45%	61.7
	S25seq	GGC TAT AGC CAA TCG GTG GTT TAT CGT G	50%	61.4
gp47	pH_S47F	GCG CCA TGG CAG GTA ACG CCA CCC A	68%	65.9
	pH_S47R	GCG TCT AGA TTA TTC GCC GTC AAG CAT ATT TTT GCG	44%	64.4
	S47seq	CAA CGA CAC CGC TGC GAC TAA CC	61%	60.6
gp70	pH_S70nxF	GCG GAA TTC ATG TAC AGC ATT ACT GAT TTT ATC AAA AG	34%	61.2
	pH_S70nxR	GCG TCT AGA CTA CGG TTT ATA ATC AAC TTC G	42%	60.4
	S70seq	GTT AAC CGA TGC GGC CAT TGT CCA G	56%	61.0
gp171	pH_S171F	GCG CCA TGG CTA CTG CTT ATC AAG TGC	56%	62.8
	pH_S171R	GCG TCT AGA TTA TTT TTT TCG TGC TTT GTA GCG ATG	39%	62.2
	S171seq	GGC CTA CCG AAG CTC GAA TTC GTT GAG	56%	62.8
gp168	S168F	GCT TTA TAC CGT ATC TGA CGC ACA GCG TCG	53%	64.4
	S168R	GGT ACG GAA CAT AAG ATT GAC CTG GAT AAC ATC AGG	44%	64.4
	S168seq	GCA TCG TTG GGA ATG GAG TGG GTA GGT	56%	62.8

gp203	pH_S203F	GCG CCA TGG CTG ATT TAC AAG GCT TTA TTC	47%	61.6
	pH_S203R	GCG TCT AGA TTA GGC TTC TGG CTC TAC	52%	61.3
	S203seq	TGA ATT CCA GCG TTT GCT GTT GGG GC	54%	61.1
gp214	pH_S214F	GCG CCA TGG ATG TTA ATC CAC TTG CAG C	54%	62.9
	pH_S214R	GCG TCT AGA TTA TTT TTT GTT TGC TTT CAG GGT CTT G	38%	62.2
	S214seq	GTC CAT CCT AAT GTG GCG GCA GAG	58%	60.8
gp219	pH_S219F	GCG CCA TGG GAT TAG GAC GCC GTT TC	62%	64.3
	pH_S219R	GCG TCT AGA TCA TGG CGT CTG CGT TTG C	57%	64.3
	S219seq	CGA AGT GGT AGG AAT TCA GGA CTG G	52%	59.3
gp238	pH_S238nxF	GCG CCA TGG AAG CAA GAC AAC TAA AAG ATT C	45%	61.7
	pH_S238nxR	GCG TCT AGA CAC TTC CGG GAA GTC AAT	52%	61.3
	S238seq	GGA TGC TCA GTC ACA AGG AGG TAA AGC	52%	61.3
gp241	pHgpS241F	GCG CCA TGG ATG CTG TGA TTA AAC AGG AAG TTG AGC	50%	66.7
	pHgpS241R	GCG TCT AGA ACA TCC TTT TCT TTG ATG TTC TGT ACC GG	45%	65.5
	S241seq	AGC AAC GTT CAA CGC AAT TGG GTA C	48%	57.7

a. Primers with names beginning with “pH” include restriction endonucleases sites, if they have “nx” they cannot be used for cloning or gene expression in the plasmid pHERD, the ones ending with “F” are forward primers, with “R” are reverse primers, and with “seq” are for Sanger sequencing.

The DNA sequencing chromatograms were visualized in Chromas Lite and were manually checked to ensure the quality of the data. The peaks were evenly-spaced and the baseline was minimal, special focus was placed on the regions where the mutations were detected to ensure there were no mis-called nucleotides.

BLASTn alignments of Sanger reads for each mutant sample against the SPN3US genome confirmed the nonsense mutations (C to T or G to A) for each amber mutant, as an example the alignment for amber mutant 1 is shown in Figure 13 and the regions where the amber mutation is present in the alignments for the mix F (amber 21 and 28) is shown in Figure 14. All wild-type control sequences were as expected. Amber mutant phages 25 and 31 that were combined in mix I showed an amber mutation in the same position in the gene encoding for gp241.

Score	Expect	Identities	Gaps	Strand
584 bits(316)	2e-168	321/323(99%)	1/323(0%)	Plus/Plus

```
Query  46298  AGGCGGTTAACAACACGATTAAAGTGTTGGAGCAGTTCCAGAAAGCAACCAAGGATAACG  46357
              |||||||| ||||||||||||||||||||| |||||||||||||||||||||||||||
Sbjct  17     AGGCGGTT-ACAACACGATTAAAGTGTTGGAGTAGTTCCAGAAAGCAACCAAGGATAACG  75

Query  46358  TTGAATACGTTAACCTGATTAGCGAAGTCATCATGAATGAGCTTAATCAGCCTTCTGGCC  46417
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  76     TTGAATACGTTAACCTGATTAGCGAAGTCATCATGAATGAGCTTAATCAGCCTTCTGGCC  135

Query  46418  TGCGTACCGACGGTCGTGAACTCTCGATGGTGTTATTCTCCCCAGACGGGGATGAAGACC  46477
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  136    TGCGTACCGACGGTCGTGAACTCTCGATGGTGTTATTCTCCCCAGACGGGGATGAAGACC  195

Query  46478  TCTTCGTCGGCTTGCGACGTAACCTGTCTTTGGTGCGTCGTTGTTGCGGTGCTGCCGCTA  46537
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  196    TCTTCGTCGGCTTGCGACGTAACCTGTCTTTGGTGCGTCGTTGTTGCGGTGCTGCCGCTA  255

Query  46538  CACTGCAGACGATACTGAACCTGGCCAAACGTAACCCGTTGGTGAAAGGCAATAATCGCA  46597
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  256    CACTGCAGACGATACTGAACCTGGCCAAACGTAACCCGTTGGTGAAAGGCAATAATCGCA  315

Query  46598  AAAATATGCTTGACGGCGAATAA  46620
              |||||||||||||||||||||||
Sbjct  316    AAAATATGCTTGACGGCGAATAA  338
```

Figure 13: Alignment of the sequenced region of the amber mutant 1 (gp47) against the reference SPN3US phage genome using BLASTn. This example illustrates a typical alignment and the quality of the Sanger reads obtained, which is a good confirmation of the previous sequencing. The mutation from C to T was confirmed at position 46330 causing a change from glutamine codon (CAG) to amber stop codon (TAG), shown in yellow. Query sequence is the reference genome and subject is the sequenced region of the amber mutant 1.

A

Wild-type

```
Query  186842  CAGTTCATTAATCTGTTTCATAAACAGTCCGAACAGCTCAACCCAGTCAGCACCCA  186897
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  72      CAGTTCATTAATCTGTTTCATAAACAGTCCGAACAGCTCAACCCAGTCAGCACCCA  17
```

Amber 21

```
                          GAC
Query  186842  CAGTTCATTAATCTGTTTCATAAACAGTCCGAACAGCTCAACCCAGTCAGCAC  186894
               ||||||||||||||| ||||||||||||||||||||||||||||||||||||
Sbjct  70      CAGTTCATTAATCTATTTCATAAACAGTCCGAACAGCTCAACCCAGTCAGCAC  18
                          GAT
```

Amber 28

```
Query  186842  CAGTTCATTAATCTGTTTCATAAACAGTCCGAACAGCTCAACCCAGTCAGCACCCA  186897
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  69      CAGTTCATTAATCTGTTTCATAAACAGTCCGAACAGCTCAACCCAGTCAGCACCCA  14
```

B

Wild-type

```
Query  203933  TCTCATCCATCGAGTTTGCTTCAGTCTGACTTTTAGCAACCTGAGCCTGGTTCTTTTCAA  203992
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  118     TCTCATCCATCGAGTTTGCTTCAGTCTGACTTTTAGCAACCTGAGCCTGGTTCTTTTCAA  59
```

Amber 21

```
Query  203931  GTTCTCATCCATCGAGTTTGCTTCAGTCTGACTTTTAGCAACCTGAGCCTGGTTCTTTTC  203990
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  115     GTTCTCATCCATCGAGTTTGCTTCAGTCTGACTTTTAGCAACCTGAGCCTGGTTCTTTTC  56
```

Amber 28

```
                                                      GAC
Query  203931  GTTCTCATCCATCGAGTTTGCTTCAGTCTGACTTTTAGCAACCTGAGCCTGGTTCTTTTC  203990
               |||||||||||||||||||||||||||||||||||||||||| ||||||||||||||||
Sbjct  116     GTTCTCATCCATCGAGTTTGCTTCAGTCTGACTTTTAGCAACCTAAGCCTGGTTCTTTTC  57
                                                      GAT
```

Figure 14: Sequenced regions of the proteins gp214 and gp238 amplified from the amber mutants 21 and 28 (Mix F) and the wild-type phage, aligned against the reference SPN3US phage genome using BLASTn. A) The first three alignments corresponding to the protein gp214 show that a mutation could be detected only in the amber mutant 21 (in yellow). B) The same can be observed for gp238 presenting an amber mutation only in amber mutant 28 (in yellow). This confirms the amber mutation sites identified from the Illumina sequencing. Both proteins gp214 and gp238, are encoded in the complementary strand, so the mutations observed from G to A in the leading strand translated to mutations from C to T in the coding strand for these proteins resulting in amber stop codons (TAG). Query sequence is the reference genome and subject is the sequenced region of the amber mutants or the wild-type phage.

6.3 Analyses aimed to detect homologs to the newly identified SPN3US essential genes

The SPN3US proteins gp260, gp256 and gp75 belong to the 20% of proteins whose putative function was annotated as they shared homology to functionally assigned proteins available at the time of the genome announcement (Lee et al. 2011). For this reason, they were chosen as control proteins as a well conserved sequence similarity among the ϕKZ-related phages had been previously observed. The control genes and their corresponding proteins were submitted to the different BLAST searches to witness the sensitivity of the tests. This also helped to determine which the best method was for finding similar sequences in our diverged phage group (Table 7). Even though these are highly conserved proteins, gp260 (terminase) was the only protein to which a domain was found in the Conserved Domain Database, which demonstrates how divergent these phages are and how difficult it is to annotate their genomes (Fig 15).

Table 7. Summary of matches to the control SPN3US genes and their protein products in other ϕKZ-related phages level using different BLAST tools. ✓ refers to a match identified, - refers to no match identified.

Phage	gp	BLASTn	tBLASTn	BLASTp	PSIBLAST
Erwinia phage ϕEaH2	260 terminase	✓	✓	✓	✓
	256 sheath	✓	✓	✓	✓
	75 major capsid	✓	✓	✓	✓
Cronobacter phage CR5	260 terminase	-	✓	✓	✓
	256 sheath	-	✓	✓	✓
	75 major capsid	✓	✓	✓	✓
Pseudomonas phage ϕKZ	260 terminase	-	✓	✓	✓
	256 sheath	-	✓	✓	✓
	75 major capsid	-	✓	✓	✓
Pseudomonas phage 201ϕ2-1	260 terminase	-	✓	✓	✓
	256 sheath	-	✓	✓	✓
	75 major capsid	-	✓	✓	✓
Pseudomonas phage ϕPA3	260 terminase	-	✓	✓	✓
	256 sheath	-	✓	✓	✓
	75 major capsid	-	✓	✓	✓
Erwinia phage Ea35-70	260 terminase	-	✓	✓	✓
	256 sheath	-	✓	✓	✓
	75 major capsid	-	✓	✓	✓
Erwinia phage ϕEaH1	260 terminase	-	✓	✓	✓
	256 sheath	-	✓	✓	✓
	75 major capsid	-	✓	✓	✓
Ralstonia phage RSL2	260 terminase	-	✓	✓	✓
	256 sheath	-	✓	✓	✓
	75 major capsid	-	-	✓	✓
Vibrio phage JM-2012	260 terminase	-	✓	✓	✓
	256 sheath	-	-	✓	✓
	75 major capsid	-	-	✓	✓

Vibrio phage VP4B	260 terminase 256 sheath 75 major capsid	- - -	✓ ✓ -	✓ ✓ ✓	✓ ✓ ✓
Pseudomonas phage OBP	260 terminase 256 sheath 75 major capsid	- - -	✓ ✓ -	✓ ✓ ✓	✓ ✓ ✓
Pseudomonas phage EL	260 terminase 256 sheath 75 major capsid	- - -	✓ ✓ -	✓ ✓ ✓	✓ ✓ ✓

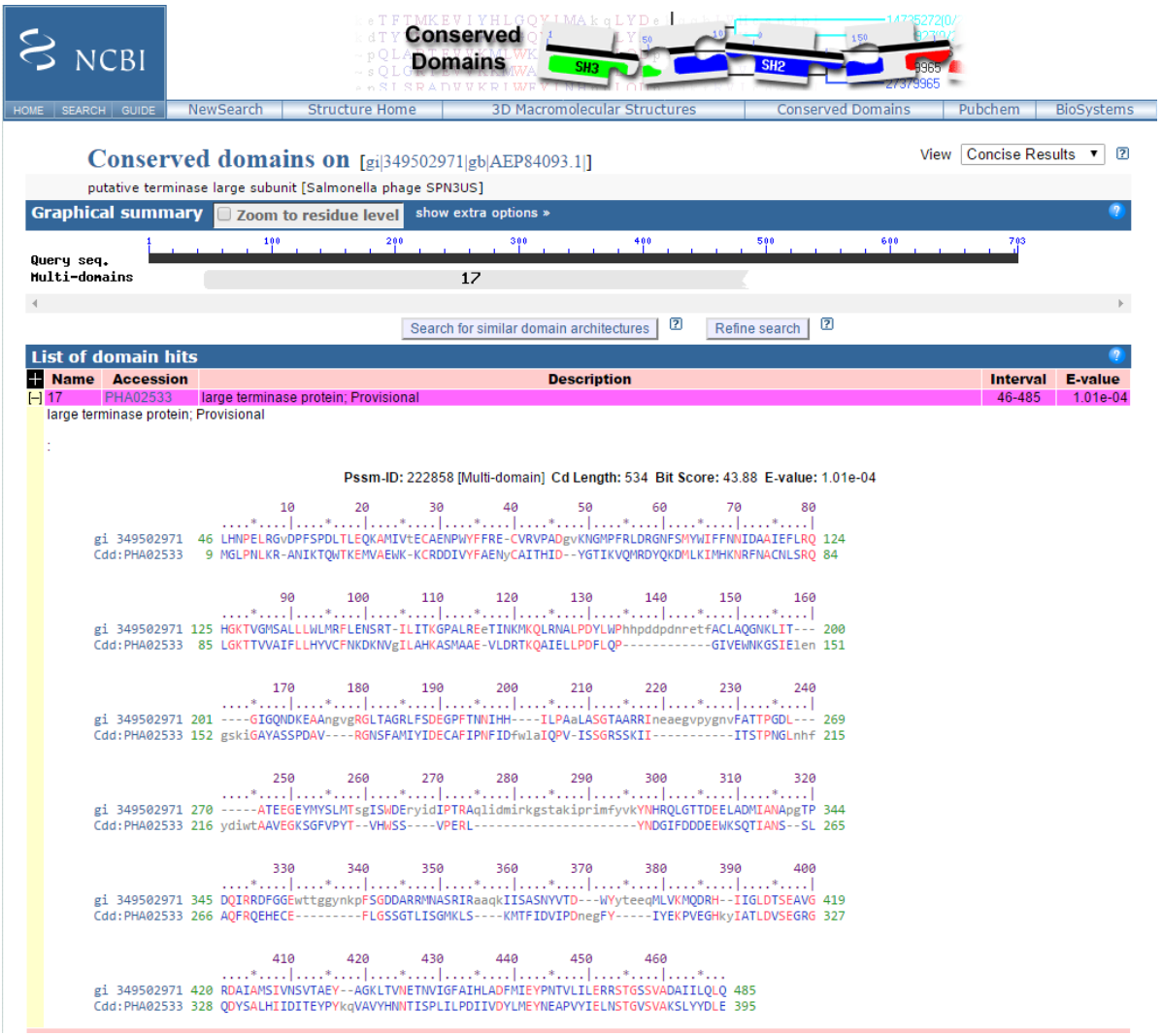

Figure 15. Conserved domain match for SPN3US gp260 terminase protein. Retrieved on October 2015 (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi)

6.3.1 Comparative results obtained between different BLAST algorithms

BLASTn searches from each of the three control genes of SPN3US (gp260, 256 and 75) resulted in the detection of similar genes in *Erwinia* phage ϕEaH2 and *Cronobacter* phage CR5 only for gp75 (Table 7 and Table 8). For only four of the eleven essential genes similar genes were found in phage ϕEaH2 (Table 8). No additional hits were found when using less stringent e-values.

Table 8. Percentage of identity of the hits found by BLASTn among the ϕKZ-related phages, of the control and newly identified SPN3US essential genes. - refers to no match identified.

	Percentage of identity (%)													
Phage	**gp 75**	**gp 256**	**gp 260**	**gp 19**	**gp 25**	**gp 47**	**gp 70**	**gp 168**	**gp 171**	**gp 203**	**gp 214**	**gp 219**	**gp 238**	**gp 241**
Erwinia phage ϕEaH2	84.92	80.16	80.03	-	-	-	-	75.08	75.57	75.41	-	-	78.21	-
Cronobacter phage CR5	76.53	-	-	-	-	-	-	-	-	-	-	-	-	-
Pseudomonas phage ϕKZ	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Pseudomonas phage 201ϕ2-1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Pseudomonas phage ϕPA3	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Erwinia phage Ea35-70	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Erwinia phage ϕEaH1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Ralstonia phage RSL2	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Vibrio phage JM-2012	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Vibrio phage VP4B	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Pseudomonas phage OBP	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Pseudomonas phage EL	-	-	-	-	-	-	-	-	-	-	-	-	-	-

The BLASTp search gave more hits than the BLASTn search. In the case of the control proteins one hit was found for each ϕKZ-related phage with the e-value cutoff of 1e-4. Using less stringent e-value cutoffs some hits had better percentage of identity but the query coverage was remarkably lower. For the essential proteins when using the e-value cutoff of 1e-4 half of the proteins had hits in all ϕKZ-related phages, the rest of the proteins either represented hits in few ϕKZ-related phages or had no hits at all which was the case for one protein (gp19). By lowering the stringency of the

e-value more hits could be found for the group of proteins that previously resulted in few hits. In order to determine if the additional hits had reliable sequence similarity the location of the new hit was considered, given that the gene encoding the protein in study was located in a conserved cluster observed only among the ϕKZ-related phages. If no synteny was observed for the gene encoding the protein being analyzed or no hits could be found in the conserved clusters, a reverse BLAST was employed with the possible hits to determine if they could have a higher similarity with other protein in phage SPN3US to discard them as good hits for the protein being analyzed.

In order to increase the number of hits, especially for the proteins that resulted in few or no hits, a tBLASTn search was employed to find similar protein coding regions in unannotated nucleotide sequences by translating the ϕKZ-related phages nucleotide sequences in all reading frames and comparing them with the identified essential proteins in SPN3US. The synteny strategy and a reverse BLAST were also performed. This search resulted in similar hits as the BLASTp search (with e-value cutoff of 1e-4) with only one possible hit for gp19 discussed below in section 4.2.2.

Overall the use of an e-value threshold of 10 did not return significant hits, for this reason the PSIBLAST search was employed with a threshold of 1. The PSIBLAST search resulted in more hits given that it is more sensitive and thus more appropriate for finding distant homologs for the translated proteins of the identified essential genes in SPN3US, and it helped to verify the identified similar proteins in the ϕKZ-related phages from the previous steps. Generally, with each iteration the percentage of identity decreased and the query coverage extended while the e-value fluctuated though tending to enhance.

The PSIBLAST search in the non-redundant database resulted in few or no hits outside the ϕKZ-related phages for the majority of the identified essential proteins. Proteins gp168, gp214, gp238 and gp241 had hits that led to further analyses, and these are detailed in the sections below.

The following sections present in more detail the results obtained from the sequence similarity analyses for each identified essential gene and their gene product in SPN3US. Two parameters were used to classify the essential genes to display the results, how well they are conserved among the ϕKZ-related phages and if they encode for virion (structural) proteins. The analyses of the amber mutant phage presenting amber mutations in two genes are consolidated in a single section (4.3.4).

6.3.2 Identified essential genes showing well conserved sequence similarity in the ϕKZ-related phages

The majority of the essential genes identified in the present research were demonstrated to be well conserved as similar sequences were found in all ϕKZ-related phages, especially in the genes encoding the virion structure (Table 9). It is important to highlight that the highest percentages of identity were observed with proteins from phages ϕEaH2 and CR5 (Table 10). Table 9 also summarizes the findings regarding the syntenic relationships of the genes and whether or not they code for a structural protein as determined by the mass spectral data from the SPN3US wild type phage.

Table 9. Summary of results of various analyses of control and newly identified SPN3US essential genes. Control proteins are shaded in blue.

GP	Well conserved	In a Synteny region	Virion protein	Processed by prohead protease
75	✓	✓	✓	✓
256	✓	✓	✓	×
260	✓	✓	✓	×
19*	×	✓	×	×
25	✓	✓	✓	×
47	×	✓	✓	✓
70	×	×	×	×
77	✓	✓	×	×
168	✓	✓	✓	×
169	✓	✓	✓	×
171	✓	✓	✓	×
203	✓	✓**	✓	×
214	×	✓**	✓	×
219	✓	✓**	×	×
220	×	✓**	×	×
238	✓	✓	✓	×
241	✓	✓	✓	×
255	✓	✓	✓	×

* Still unclear if it is essential or not

** Present in the SPN3US-like phages

Table 10. Summary of the hits found by PSIBLAST among the ϕKZ-related phages, of the proteins encoded by the identified essential genes of SPN3US found in amber mutant phages. The values shown correspond to the percentage of identity. - refers to no match identified.

	Percentage of identity (%)													
Phage	**gp19**	**gp25**	**gp47**	**gp70**	**gp77**	**gp168**	**gp171**	**gp203**	**gp214**	**gp219**	**gp220**	**gp238**	**gp241**	**gp255**
Erwinia phage ϕEaH2	-	69.05	56.34	29.24	82.26	43.54	80.95	80.49	78.09	64.50	-	75.03	78.29	90.38
Cronobacter phage CR5	-	51.24	26.82	-	57.50	31.4	46.43	56.86	47.22	44.05	39.19	53.22	58.84	71.72
Pseudomonas phage ϕKZ	-	20.00	-	-	20.71	16.16	19.76	29.78	19.72	23.87	21.90	22.49	28.78	26.28
Pseudomonas phage 201ϕ2-1	-	21.43	-	-	17.89	10.27	22.60	29.05	14.08	22.50	20.62	21.43	32.46	24.91
Pseudomonas phage ϕPA3	-	15.29	-	-	19.66	14.93	20.23	29.80	14.35	22.31	13.25	20.90	38.87	26.28
Erwinia phage Ea35-70	-	24.22	25.00	-	16.89	15.84	21.72	24.67	-	22.03	-	22.40	28.52	27.41
Erwinia phage ϕEaH1	-	16.67	26.09	-	20.20	14.74	19.53	28.60	19.53	17.03	26.88	23.60	27.03	29.35
Ralstonia phage RSL2	-	17.69	-	-	19.17	16.28	16.42	24.28	13.27	22.22	24.19	25.39	26.63	31.08
Vibrio phage JM-2012	-	16.95	-	-	19.90	14.21	-	23.11	17.44	22.03	-	20.33	24.88	21.25
Vibrio phage VP4B	-	19.69	-	-	15.81	11.87	19.65	24.19	-	17.55	-	17.31	25.35	20.16
Pseudomonas phage OBP	-	21.71	-	-	15.95	13.18	18.93	23.54	-	17.92	-	19.53	23.48	17.87
Pseudomonas phage EL	-	25.60	-	-	15.22	12.62	17.63	23.62	-	17.89	-	18.47	28.77	19.77

6.3.2.1 Well conserved essential genes encoding for virion proteins.

Mass spectral data from the wild type phage showed that 88 genes encode for the proteins that are found associated with the virion, corresponding to 46% (111 kbp) of the genome. A conserved protease cleavage motif, AXE-cleavage (where X represents any amino acid), was found in the SPN3US prohead proteins involved in the morphogenesis of the capsid (Thomas unpublished).

The gene encoding for the structural **protein gp25** is located in a region that has conservation of gene order, especially in the SPN3US-like phages as can be seen in the genome diagram elaborated for gp19 in section 4.3.4. As stated before, PSIBLAST returned more relevant hits than the other BLAST tools, helping to find similar proteins in all ϕKZ-related phages (Table 11); however, no homologs were found outside the ϕKZ-related subfamily. The percentage of identity of the proteins with similar sequence to gp25 are shown in Table 10.

Protein **gp169** demonstrates to be well conserved among the ϕKZ-related phages. The analyses of this protein are included in the section 4.3.4 as it is a paralogous protein of gp168.

The gene encoding **protein gp171** is located in a syntenic region as can be seen in the genome diagram elaborated for gp168 in section 4.3.4. It has homology at the nucleotide level with phage ϕEaH2 and is conserved in all ϕKZ-related phages except for *Vibrio* phage JM-2012 (Table 10 and Appendix B.3 and C.6). No hits were found outside the ϕKZ-related phages.

Table 11. Presence or absence of BLAST matches for the conserved protein gp25 among the ϕKZ-related phages. This table is provided as an example, other summaries of matches for essential genes can be found in appendix B. - refers to no match identified.

Phage	**BLASTn**	**tBLASTn**	**BLASTp**	**PSIBLAST**
Erwinia phage ϕEaH2	-	✓	✓	✓
Cronobacter phage CR5	-	✓	✓	✓
Pseudomonas phage ϕKZ	-	✓	✓	✓
Pseudomonas phage 201ϕ2-1	-	-	-	✓
Pseudomonas phage ϕPA3	-	✓*	-	✓
Erwinia phage Ea35-70	-	✓	✓	✓
Erwinia phage ϕEaH1	-	-	-	✓
Ralstonia phage RSL2	-	-	-	✓
Vibrio phage JM-2012	-	-	-	✓
Vibrio phage VP4B	-	✓*	✓*	✓
Pseudomonas phage OBP	-	✓*	✓*	✓
Pseudomonas phage EL	-	-	✓*	✓

* Defined by synteny analysis in hits with higher e-values

The gene encoding the structural **protein gp203** has homology at the nucleotide level with phage ϕEaH2, is very well conserved in all ϕKZ-related phages (Table 10) with hits by BLASTp, PSIBLAST, and even in tBLASTn (Appendix B.4). However, the gene order of the region where gp203 is located is not conserved in all the subfamily phages and synteny can be seen only among the phages that belong to the genus. No hits were found in the non-redundant database outside the ϕKZ-related phages.

The genes encoding the structural **proteins gp238 and gp241** are located in a region with well conserved gene order in all the subfamily phages, an inversion of the gene cluster is observed in the ϕKZ-like phages. The three BLAST tools using a protein query returned hits for all ϕKZ-related phages for both proteins but only gp238 presented homology at the nucleotide level with phage ϕEaH2 (Table 8 and Table 10, Appendix B.7, B.8, C.11 and C.12).

One phage outside the ϕKZ-related phages, *Yersinia* phage ϕR1-37, resulted as a hit in the PSIBLAST search for both proteins gp238 and gp241. Gp238 had several hits with proteins in bacteria related to phage infection, these results were maintained in the sixth iteration of PSIBLAST. From the hits related to the phage infection protein (yhgE/Pip) the one with the best e-value (Multispecies: YhgE/Pip domain-containing protein [*Bacillus cereus* group] with accession number WP_016126510.1) was selected to perform a reverse BLAST, the search did not return hits for any ϕKZ-related phage. Five conserved domain hits were detected for this protein (Fig 16). A prediction of transmembrane helices in for gp238 was performed using the TMHMM2.0 online server (Sonnhammer et al. 1998). We obtained one predicted transmembrane helix with 32 expected amino acids which suggests that it is likely to be a protein with a function associated with host take over (Fig 17).

The PSIBLAST search for gp241, resulted in numerous hits related to DNA-directed RNA polymerase subunits. Gp241 was one of the proteins with an already assigned function, and was the only protein from the newly identified essential proteins to have a conserved domain in the CDS (Fig 18).

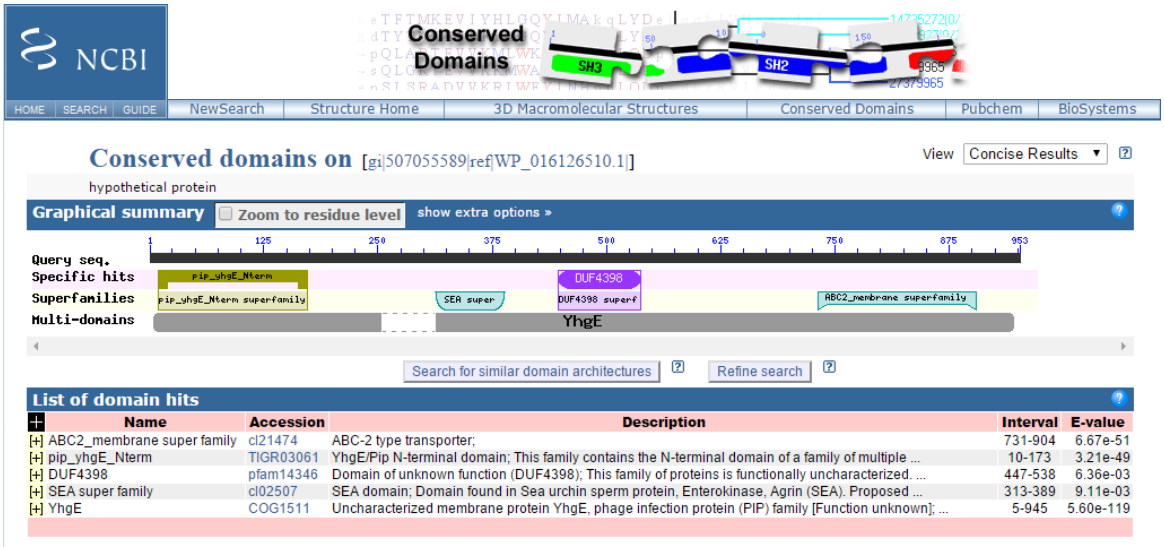

Figure 16. Conserved domain found for protein yghE/Pip of *Bacillus cereus* group, which was determined to be similar to SPN3US gp238 in a PSIBLAST search of the non-redundant databases (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi)

```
# gi238|349502949|gb|AEP84071.1| Length: 740
# gi238|349502949|gb|AEP84071.1| Number of predicted TMHs:  1
# gi238|349502949|gb|AEP84071.1| Exp number of AAs in TMHs: 32.10509
# gi238|349502949|gb|AEP84071.1| Exp number, first 60 AAs:  0.00526
# gi238|349502949|gb|AEP84071.1| Total prob of N-in:        0.67861
gi238|349502949|gb|AEP84071.1|  TMHMM2.0        inside       1   151
gi238|349502949|gb|AEP84071.1|  TMHMM2.0        TMhelix    152   174
gi238|349502949|gb|AEP84071.1|  TMHMM2.0        outside    175   740
```

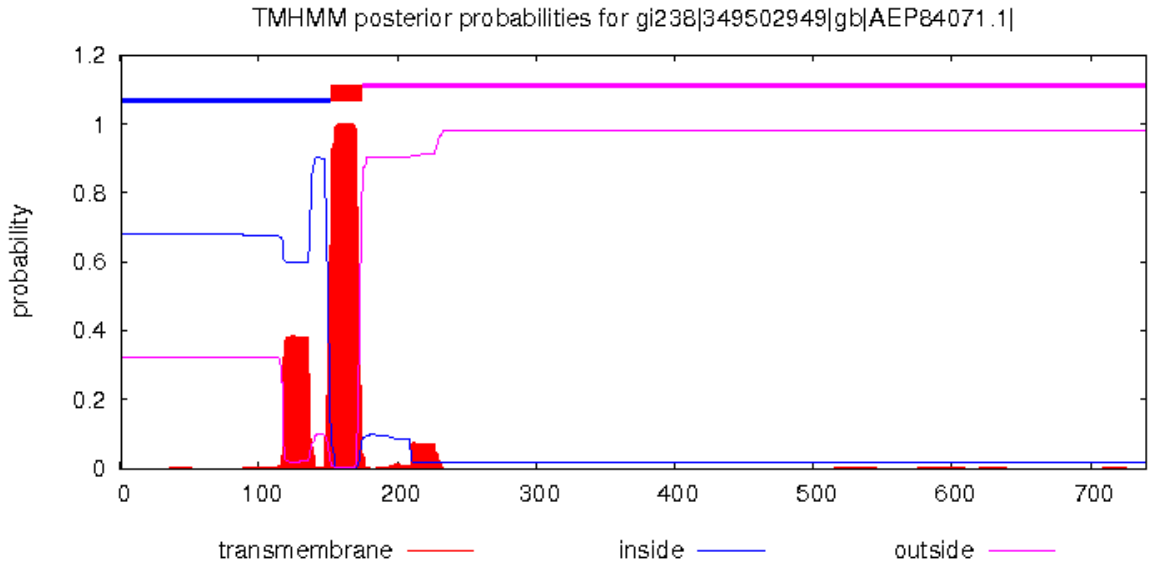

Figure 17. Prediction of transmembrane helices for SPN3US gp238 using TMHMM. (http://www.cbs.dtu.dk/services/TMHMM/)

Figure 18. Conserved domain found for SPN3US gp241. Retrieved on October 2015. (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi)

The gene encoding for the protein **gp255** was sent in the last group of samples sequenced by MiSeq and was submitted to a PSIBLAST search only (Appendix C.13). It was found to be well conserved among the ϕKZ-related phages and is located in a syntenous region. Function for essential gene gp255 is known as it has homology to gp30 in ϕKZ which is related to tail tube.

6.3.2.2 Well conserved genes encoding non-structural proteins

The gene encoding **protein gp219** was determined to be conserved in all ϕKZ-related phages based on the hits of the PSIBLAST search (Table 10, Appendix B.6 and C.9), though no hits were found outside the ϕKZ-related phages. Conservation of gene order was observed among the SPN3US-like and ϕKZ-like phages.

The gene encoding for the protein **gp77** was sent in the last group of samples sequenced by MiSeq. The PSIBLAST search demonstrated it is conserved in all ϕKZ-related phages (Appendix C.5) and it is located in a syntenic region. No further analyses were performed for gp77 as it was already known that it is a non-virion RNAP subunit as demonstrated by Ceyssens (2014).

6.3.3 Identified essential genes in SPN3US with less conserved similarity in the ϕKZ-related phages

6.3.3.1 Less conserved genes encoding for virion proteins

The PSIBLAST search for **protein gp47** returned hits only among the SPN3US-like phages (Table 10 and Table 12). BLASTp was used to find hits for proteins gp44 to gp50. Similar proteins were found for gp44, gp45, gp48 and gp49 in each ϕKZ-related phage and a conservation in gene order could be observed in all the related phages (Fig 19). Gp285 in phage Ea35-70 and gp232 in phage ϕEaH1 were identified as having similarity to SPN3US gp47 using less stringent e-values. Both genes are located inside the syntenous region of their respective phages (Fig 19 and Appendix C.3).

Proteins found inside the syntenous regions in each ϕKZ-related phage were considered possible candidates as being diverged homologs to SPN3US gp47. The candidate proteins were submitted to reverse BLAST to confirm if they possessed higher similarity to other protein in SPN3US, and sequence similarity was observed between the candidates separating them into three groups (Table 13). From the candidate proteins gp86 in phage ϕKZ and gp148 in phage 201ϕ2-1 had been identified as head proteins by mass spectrometry and had been determined to be processed by the prohead protease (Thomas et al. 2008; Thomas et al. 2012). A similar situation was found with gp47 in SPN3US when analyzing the mass spectral results (Fig 20).

Table 12. Hits for SPN3US gp47 at the nucleotide and protein level using different BLAST tools. This table is an example of a less conserved protein among the ϕKZ-related phages, the results for the rest of the identified essential proteins can be found in the appendix B section. - refers to no match identified.

Phage	**BLASTn**	**tBLASTn**	**BLASTp**	**PSIBLAST**
Erwinia phage ϕEaH2	-	✓	✓	✓
Cronobacter phage CR5	-	✓	✓	✓
Pseudomonas phage ϕKZ	-	-	-	-
Pseudomonas phage 201ϕ2-1	-	-	-	-
Pseudomonas phage ϕPA3	-	-	-	-
Erwinia phage Ea35-70	-	-	✓*	-
Erwinia phage ϕEaH1	-	-	✓*	-
Ralstonia phage RSL2	-	-	-	-
Vibrio phage JM-2012	-	-	-	-
Vibrio phage VP4B	-	-	-	-
Pseudomonas phage OBP	-	-	-	-
Pseudomonas phage EL	-	-	-	-

* Defined by synteny analysis in hits with higher e-values

Table 13. Sequence similarity found using BLASTp among the proteins encoded by the genes contained in the syntenous regions detected in each ϕKZ-related phage, which demonstrated similarity with the SPN3US genes (from gp44 to gp50).

Phage	Locus tag	Accession number	aa length	similar sequences
Pseudomonas phage ϕKZ	gp86	NP_803652	428	gp89 in ϕPA3 and gp148 in 201ϕ2-1
Pseudomonas phage 201ϕ2-1	gp148	YP_001956872	414	gp86 in ϕKZ and gp89 in ϕPA3
Pseudomonas phage ϕPA3	gp89	AEH03513	413	gp148 in 201ϕ2-1 and gp86 in ϕKZ
Erwinia phage Ea35-70	gp285	YP_009005081.1	418	gp232 in ϕEaH1
Erwinia phage ϕEaH1	gp232	YP_009010285.1	456	gp285 in Ea35-70
Ralstonia phage RSL2	gp176	BAQ02704	596	
Vibrio phage JM-2012	gp20	YP_006383300	506	
Vibrio phage VP4B		AGB07224	474	gp102 in OBP and gp54 in EL
Pseudomonas phage OBP	gp102	YP_004958009	527	AGB07221 in VP4B and gp54 in EL
Pseudomonas phage EL	gp54	YP_418087	505	AGB07224 in VP4B and gp102 in OBP

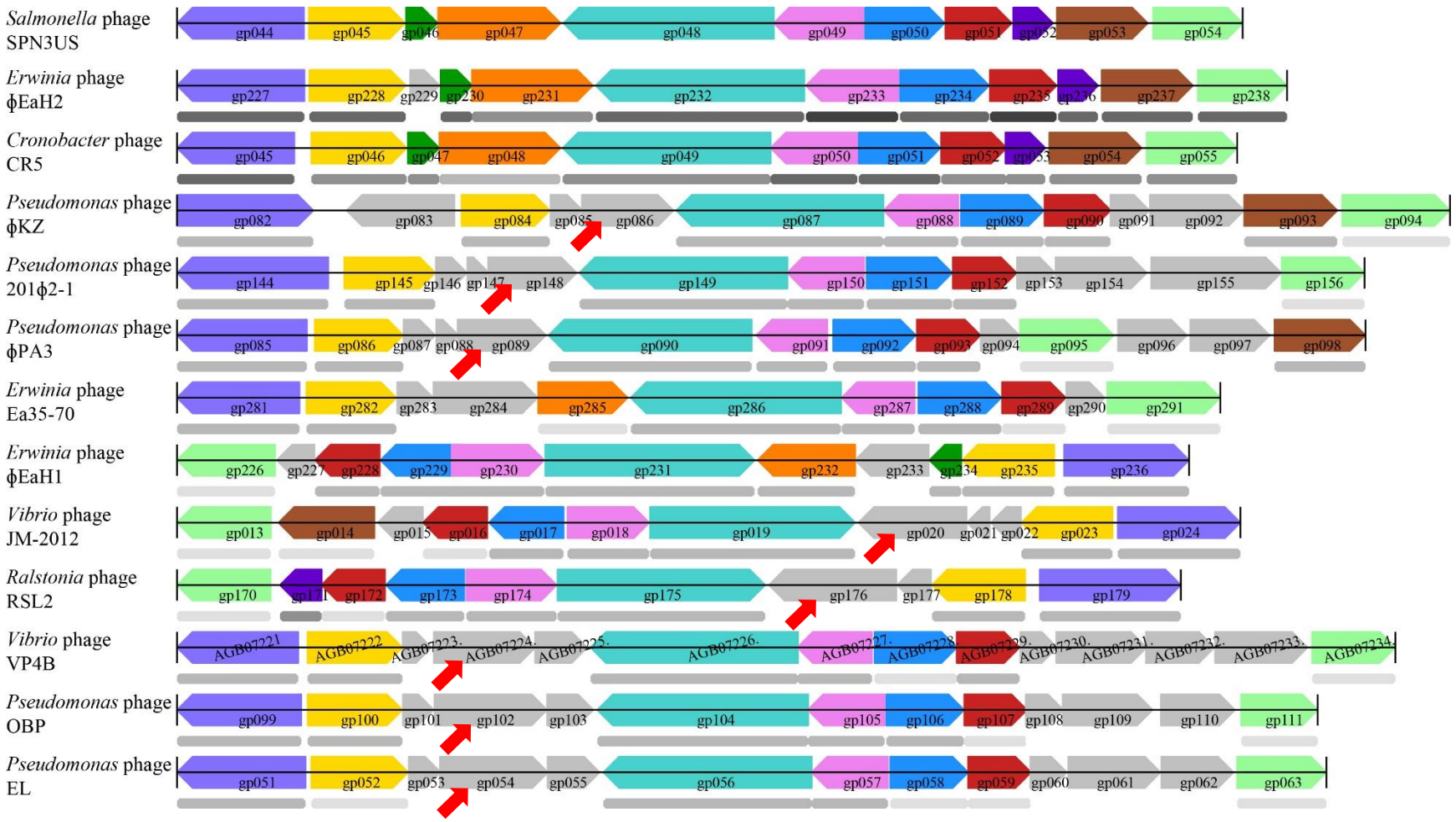

Figure 19: Comparative genomic map of the region flanking the gene coding for protein gp47 of SPN3US and the corresponding syntenic regions in other ϕKZ-related phages. Each arrow box indicates the transcriptional orientation, contains the orf number and is colored according to the protein similarity with SPN3US. The gray areas below each box indicate the degree of similarity at the protein level by shade intensity. Gp47 is shown in orange and a red arrow indicates the candidate proteins we suggest are distant homologs of gp47.

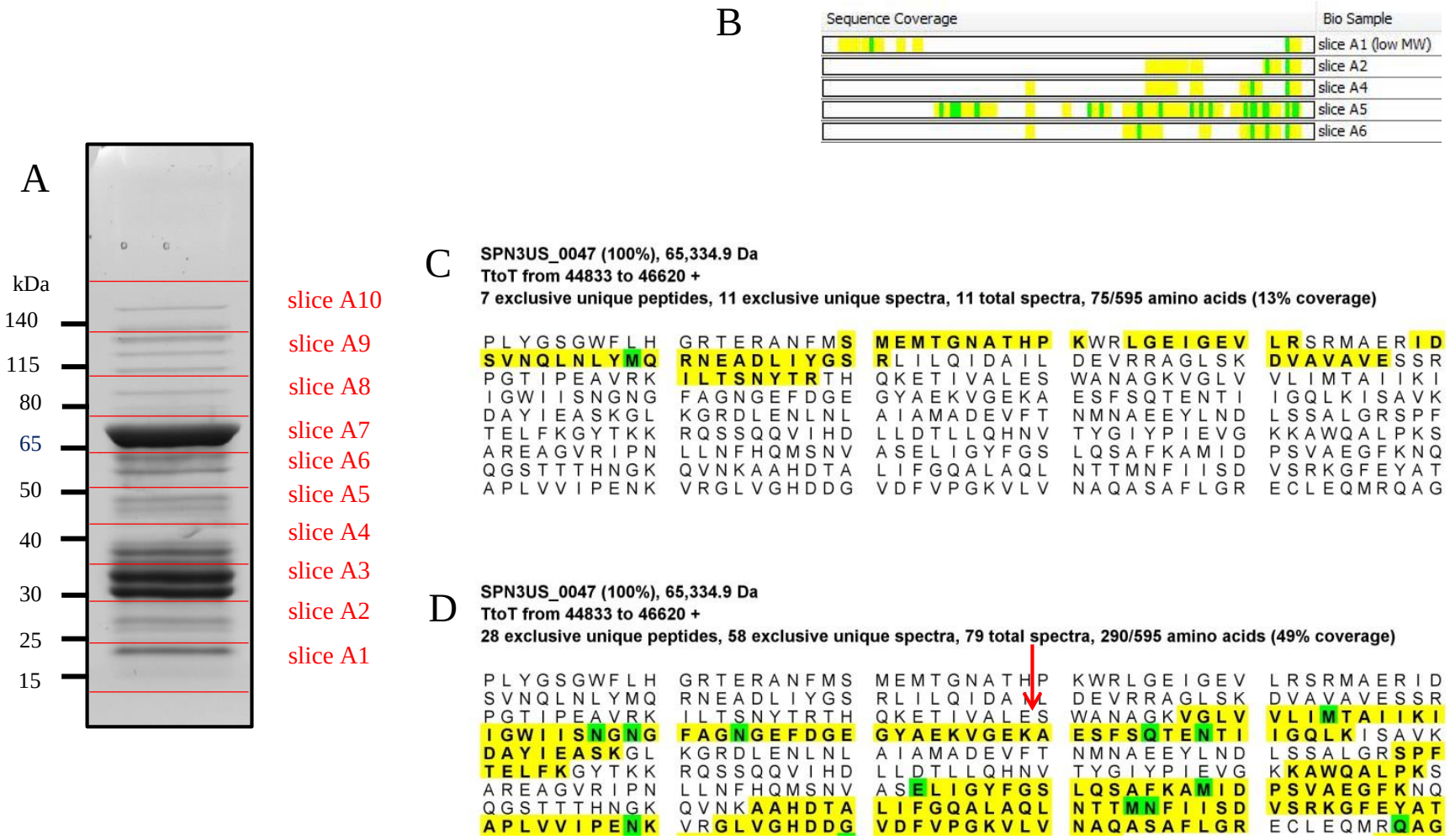

Figure 20. SPN3US is a virion protein determined to be a head protein as it is processed by the prohead morphogenic protease. A) Slices excised from an SDS-PAGE gel for mass spectrometry identification of capsid proteins, B) Shows the mass spectral peptide coverage per slice (yellow), C and D show the mass spectral peptide coverage for gp47, C) Slice A1, semi-tryptic peptide, D) Slice A5, the red arrow shows the cleavage site (ALE-residue) identified by mass spectrometry.

The gene encoding **protein gp214** is not well conserved among the ϕKZ-related phages at the sequence level (Table 10, Appendix B.5 and C.8). Conservation of gene order was observed among the SPN3US-like phages for the region containing gp214.

The PSIBLAST search in the NCBI non-redundant database resulted in no similar sequences in phages outside the ϕKZ-related phages. Several hits were found for bacteria with annotated functions related to chemotaxis in the third and fourth iterations. In order to ensure that the hits would not be discarded in future iterations due to lack of conservation two more iterations were performed. The chemotaxis related hits were discarded.

6.3.3.2 Less conserved genes encoding for non-virion proteins

BLAST searches of the gene encoding **protein gp70** and the gene product resulted in a single hit in phage ϕEaH2 (Table 10). A few hits with less conserved e-values were found when using the tBLASTn tool, though when doing a reverse BLASTp a higher similarity was found to other proteins in SPN3US so these hits were discarded. There was no conservation in the order of the genes right upstream and downstream of the gene encoding the protein gp70.

The gene encoding for the protein **gp220** was sent in the last group of samples sequenced by MiSeq. The non-structural protein gp220 is encoded by a gene that is not well conserved among the ϕKZ-related phages (Appendix C.10). Synteny can be observed only among the SPN3US-like phages; however, within the genus gp220 was only found in CR5.

6.3.4 Analysis of the double mutation in the mutant phage, amber 26

Amber mutant 26 presented an amber mutation in the genes encoding both gp19 and gp168. In order to determine if both gp19 and 168 are essential or only one of them is essential more in depth analyses were performed.

Gp168 encodes for a large structural protein (1727 aa) and was one of the few genes that demonstrated similarity at the nucleotide level with one of the ϕKZ-related phages, *Erwinia* phage ϕEaH2 (Table 8) which suggests a strong similarity. The results in PSIBLAST were very similar to the ones in BLASTp, they showed that it belongs to a group of paralog proteins in SPN3US:

gp167, gp168, gp169 and gp170, which suggests that they all have the same specialized function. The genes encoding these four proteins had a strong syntenous relationship with four proteins of similar length both in phage ϕEaH2 and phage CR5 (Table 14). Similar proteins could also be observed in the other members of the ϕKZ-related phages (Table 15). For example, in the ϕKZ-like phages (ϕKZ, 201ϕ2-1, ϕPA3) a group of five paralogs were reported and for one of the homologs in ϕKZ, gp131, Sycheva et al. (2012) has a defined crystal structure of the C-terminal domain (PDB 4GBF). They also reported that gp131 is a tail protein located in the periphery of the baseplate and is probably associated with the tail fibers (Sycheva et al. 2012). The alignment of gp131 of ϕKZ with gp168 of SPN3US showed that the similarity found is in the N-terminal domain (Fig 21), which in general has been observed with the other ϕKZ-related phages (Cornelissen et al. 2012), denoting that this region is well conserved relative to other phages (Fig 22). The secondary structure was predicted for the homologous proteins of gp168, the N-terminal domain (aa 1- aa 840) had a high alpha helical content while the C-terminal domain (aa 841 – aa 1727) had mostly beta sheet structures (Appendix D.1 and D.2).

Table 14. Summary of PSIBLAST matches for the paralogous proteins of gp168 in the phages most closely related to SPN3US

Phage	**Accession number**	**Title**	**length (aa)**	**% of id.**	**coverage**	**E-value**	**Bitscore**
Salmonella phage SPN3US	AEP84001.1	hypothetical protein SPN3US_0168	1727	100	100	0	1367
	AEP84000.1	putative virion structural protein 20	400	16.51	12	3.00E-31	139
	AEP84002.1	hypothetical protein SPN3US_0169	1376	19.87	34	2.00E-78	297
	AEP84003.1	hypothetical protein SPN3US_0170	1237	18.64	42	6.00E-05	58.5
	AEP84003.1	hypothetical protein SPN3US_0170		16.52	42	4.00E-53	215
	AEP84003.1	hypothetical protein SPN3US_0170		20.11	42	1.00E-21	113
Erwinia phage ϕEaH2	YP_007237740.1	hypothetical protein	1676	43.59	100	0	1103
	YP_007237739.1	putative virion structural protein	399	16.26	19	2.00E-32	143
	YP_007237741.1	hypothetical protein	1389	16.62	38	8.00E-75	285

	YP_007237742.1	hypothetical protein	1236	12.16	57	8.00E-05	58.1
	YP_007237742.1	hypothetical protein		17.67	57	1.00E-49	204
	YP_007237742.1	hypothetical protein		21.63	57	5.00E-27	130
Cronobacter phage CR5	YP_008125899.1	hypothetical protein CR5_159	1900	30.81	84	0	770
	YP_008125898.1	putative virion structural protein 15	399	15.54	16	2.00E-33	146
	YP_008125900.1	hypothetical protein CR5_160	1340	17.81	31	2.00E-68	265
	YP_008125901.1	hypothetical protein CR5_161	1208	14.59	66	0.024	50
	YP_008125901.1	hypothetical protein CR5_161		14.63	66	7.00E-49	202
	YP_008125901.1	hypothetical protein CR5_161		22.93	66	3.00E-47	196

Table 15. Paralogous proteins present in each ϕKZ-related phage that are homologous to SPN3US paralog family gps167 – 170.

Phage	**Paralogs (gene product – gp)**				
Salmonella phage SPN3US	167	168	169	170	
Erwinia phage ϕEaH2	89	90	91	92	
Cronobacter phage CR5	158	159	160	161	
Pseudomonas phage ϕKZ	131	132	133	134	135
Pseudomonas phage 201ϕ2-1	216	217	218	219	220
Pseudomonas phage ϕPA3	150	151	152	153	154
Erwinia phage Ea35-70	17	18	19	20	
Erwinia phage ϕEaH1	189	191	192	193	
Ralstonia phage RSL2	64	65			
Vibrio phage JM-2012	55	56			
Vibrio phage VP4B*	AGB07276.1	AGB07277.1	AGB07278.1		
Pseudomonas phage OBP	142	143	144	145	
Pseudomonas phage EL	113	114	115	116	

* Accession number present in table due to lack of locus tag annotated for this genome.

Outside the ϕKZ-related phages the PSIBLAST search resulted in hits with *Escherichia* and *Enterobacteria* phages in which a similar protein has been annotated as a tail fiber protein. An alignment of gp168 with gpS of *Enterobacteria* phage P1 showed a 12.99% of identity and 42% of query coverage. P1 gpS is described as consisting of a constant N-terminal and a variable C-

terminal. Gp168 aligned with the conserved N-terminal domain of P1 gpS as well as with other similar proteins from *Escherichia* and *Enterobacteria* phages. This trend shows some conservation of this protein out of the ϕKZ-related phages. P1 gpS belongs to the conserved domain protein family pfam03406: Phage_fiber_2. The secondary structure of P1 gpS showed a similar trend as the homologs for gp168 (Appendix D.3). HHPRED (Söding 2005) was used to detect similarities for the whole gp131 and gp168 proteins, and also for the N-terminal and C-terminal domains of each protein. For gp131 in ϕKZ we found weak matches to gp10 and gp12 of the T4 phage when using the whole protein and the N-terminal domain respectively. We also found a weak match to the putative long tail fiber protein of *Listeria* phage A118 on both searches along with other matches with lower scores. The search using the whole protein and the C-terminal of ϕKZ gp131 as a query returned a single phage related hit with 4GBF, the crystal structure of the C-terminal domain. For gp168 in SPN3US we have weak and short alignments for gp12 in T4 in both whole and N-terminal searches and with phage tail repeat like proteins present in different species of bacteria. Gp10, gp11 and gp12 of T4 are paralogs of one another and form the periphery of the baseplate (Leiman et al. 2006). The outputs from the HHPRED server can be found in the appendix E.

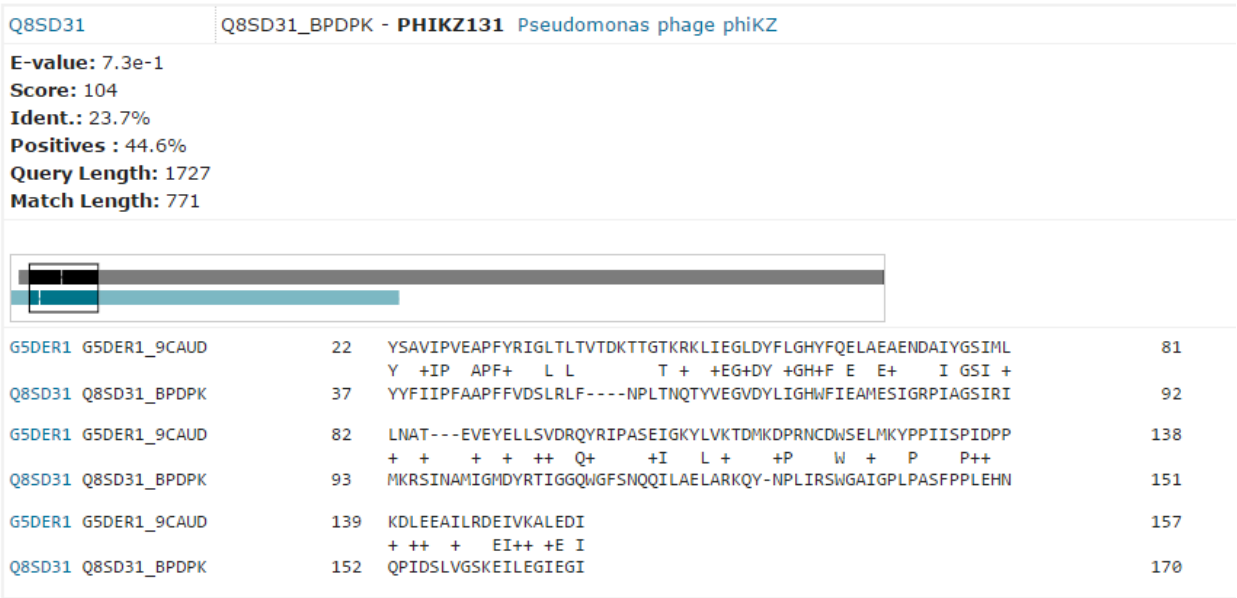

Figure 21: Alignment of gp168 of SPN3US (UniProt identifier G5DER1) against gp131 of ϕKZ (UniProt identifier Q8SD31) using the BLAST tool in UniProt (http://www.uniprot.org/blast/)


Salmonella phage SPN3US
gp166
gp167
gp168
gp169
gp170
gp171
gp172
gp173
gp174
Erwinia phage ϕEaH2
gp089
gp090
gp091
gp092
gp093
gp094
gp095
gp096
Cronobacter phage CR5
gp158
gp159
gp160
gp161
gp162
gp163
gp164
Pseudomonas phage ϕKZ
gp130
gp131
gp132
gp133
gp134
gp135
Pseudomonas phage 201ϕ2-1
gp215
gp216
gp217
gp218
gp219
gp220
Pseudomonas phage ϕPA3
gp149
gp150
gp151
gp152
gp153
gp154
Erwinia phage Ea35-70
gp016
gp017
gp018
gp019
gp020
gp021
Erwinia phage ϕEaH1
gp188
gp189
gp190
gp191
gp192
gp193
gp194
Vibrio phage JM-2012
gp055
gp056
Ralstonia phage RSL2
gp062
gp063
gp064
gp065
Vibrio phage VP4B
AGB07275.1
AGB07276.1
AGB07277.1
AGB07278.1
AGB07279.1
AGB07280.1
Pseudomonas phage OBP
gp141
gp142
gp143
gp144
gp145
gp146
Pseudomonas phage EL
gp112
gp113
gp114
gp115
gp116
gp117

Figure 22: Comparative genomic map of the region flanking the gene coding for protein gp168 of SPN3US and the corresponding syntenic regions in other ϕKZ-related phages. Each arrow box indicates the transcriptional orientation, contains the orf number and is colored according to the protein similarity with SPN3US. The gray bars below each box indicate the degree of similarity at the protein level by shade intensity and the area of alignment with the corresponding SPN3US proteins. Gp168 is shown in dark green. The green bars below gp168 show the area of alignment with the homologous proteins in the ϕKZ-related phages, the majority aligning in the N terminal domain. The bars above the SPN3US paralogs (gps167, 168, 169 and 170) indicate the area of alignment with regard to gp168.

Gp19 on the other hand encodes for a small protein (93 aa) and resulted in no hits on BLASTn, BLASTp or PSIBLAST among the ϕKZ-related phages and the non-redundant database. A conservation of the gene order was observed mainly among the SPN3US-like phages (Fig 23). Gp19 is located downstream to gp18, both in the same orientation. Gp18 is one or the two DNA polymerase subunits, the other is gp44. We confirmed that gp19 is not an extension of the DNA polymerase as the subunit gp18 covers the complete C-terminal domain.

The secondary structure of the proteins encoded by small genes downstream from the DNA polymerase detected in the other ϕKZ-related phages were compared with the secondary structure of gp19. No similarity was detected at the structural level. A similarity was found with the intergenic region between gp200 and gp201 in phage ϕEaH2 using a tBLASTn search with a 32% of identity and an e-value of 0.022. At this point we could not eliminate gp19 as being essential, but we consider it is not a well conserved gene. More analyses should be performed for gp19.

At the time of writing an amber mutant phage (115) had just been isolated and sequenced by the Phage Lab team with a single mutation in gp168, confirming that this protein is essential for SPN3US. Future work needs to be done with gp19 to reliably confirm that it is not an essential protein.

Table 16 summarizes the main findings from the present study regarding conservation among the ϕKZ-related phages and protein characteristics and functions.

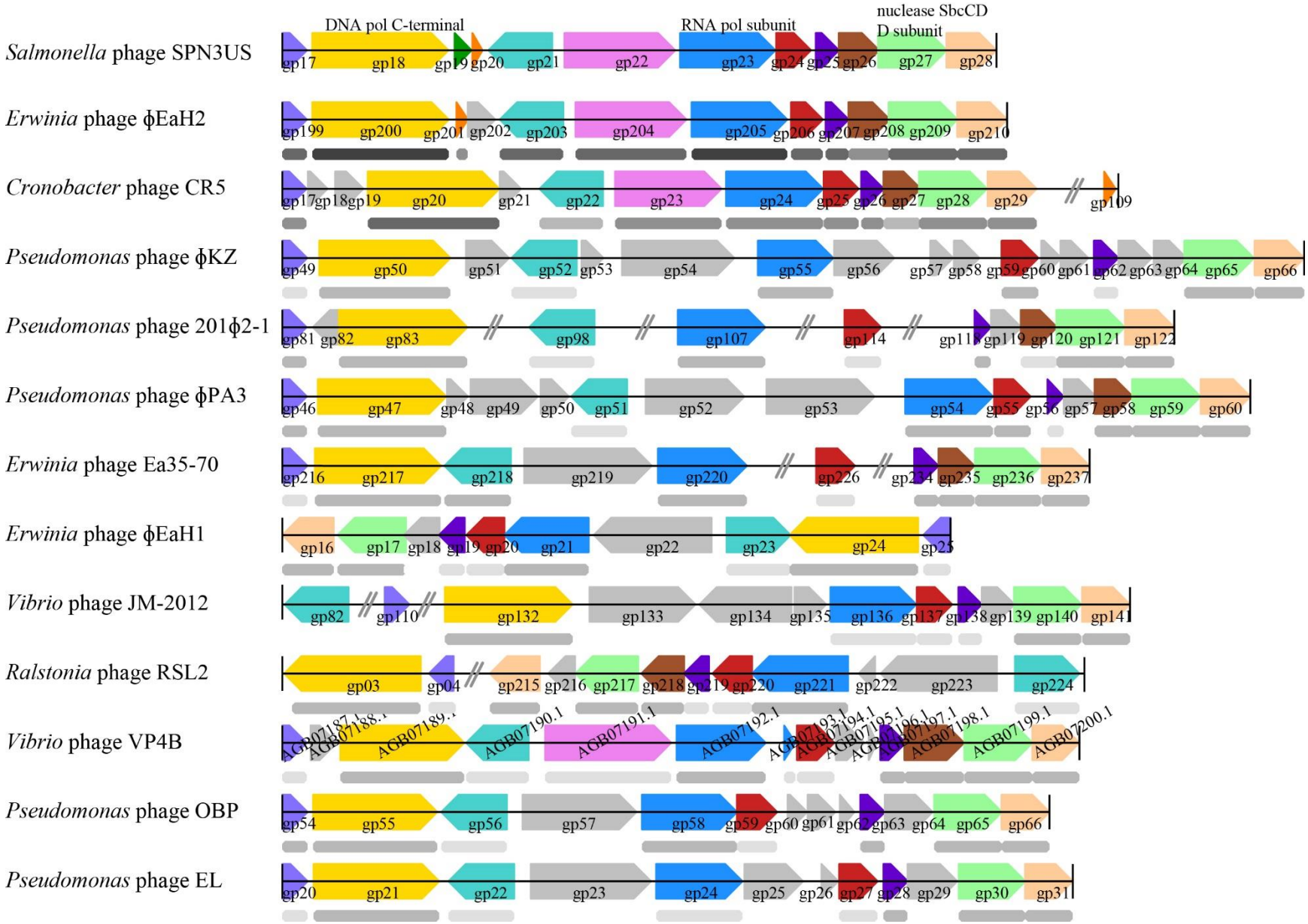
Salmonella phage SPN3US
DNA pol C-terminal
RNA pol subunit
nuclease SbcCD D subunit
gp17 gp18 gp19 gp20 gp21 gp22 gp23 gp24 gp25 gp26 gp27 gp28
Erwinia phage ϕEaH2
gp199 gp200 gp201 gp202 gp203 gp204 gp205 gp206 gp207 gp208 gp209 gp210
Cronobacter phage CR5
gp17 gp18 gp19 gp20 gp21 gp22 gp23 gp24 gp25 gp26 gp27 gp28 gp29 gp109
Pseudomonas phage ϕKZ
gp49 gp50 gp51 gp52 gp53 gp54 gp55 gp56 gp57 gp58 gp59 gp60 gp61 gp62 gp63 gp64 gp65 gp66
Pseudomonas phage 201ϕ2-1
gp81 gp82 gp83 gp98 gp107 gp114 gp119 gp120 gp121 gp122
Pseudomonas phage ϕPA3
gp46 gp47 gp48 gp49 gp50 gp51 gp52 gp53 gp54 gp55 gp56 gp57 gp58 gp59 gp60
Erwinia phage Ea35-70
gp216 gp217 gp218 gp219 gp220 gp226 gp234 gp235 gp236 gp237
Erwinia phage ϕEaH1
gp16 gp17 gp18 gp19 gp20 gp21 gp22 gp23 gp24 gp25
Vibrio phage JM-2012
gp82 gp110 gp132 gp133 gp134 gp135 gp136 gp137 gp138 gp139 gp140 gp141
Ralstonia phage RSL2
gp03 gp04 gp215 gp216 gp217 gp218 gp219 gp220 gp221 gp222 gp223 gp224
Vibrio phage VP4B
AGB07187.1 AGB07188.1 AGB07189.1 AGB07190.1 AGB07191.1 AGB07192.1 AGB07193.1 AGB07194.1 AGB07195.1 AGB07196.1 AGB07197.1 AGB07198.1 AGB07199.1 AGB07200.1
Pseudomonas phage OBP
gp54 gp55 gp56 gp57 gp58 gp59 gp60 gp61 gp62 gp63 gp64 gp65 gp66
Pseudomonas phage EL
gp20 gp21 gp22 gp23 gp24 gp25 gp26 gp27 gp28 gp29 gp30 gp31

Figure 23: Comparative genomic map of the region flanking the gene coding for protein gp19 of SPN3US and the corresponding syntenic regions in other ϕKZ-related phages. Each arrow box indicates the transcriptional orientation, contains the orf number and is colored according to the protein similarity with SPN3US. The gray areas below each box indicate the degree of similarity at the protein level by shade intensity. No hits were found for gp19 (shown in dark green) in the other phages. A conservation of gene order could be observed is clearer for the SPN3US-like phages.

Table 16. Mutant phages with the corresponding protein displaying the amber mutation, the GenBank annotated function, and new inferred functions.

Amber mutant phage	**GP**	**Annotated Function in GenBank**	**Inferred function by previous studies**	**Newly identified features in this study**	**Suggested core protein**
26	19	Hypothetical protein		Non-structural protein	Essentiality cannot be confirmed
6 / 20	25	Hypothetical protein		Structural protein	Yes
1	47	Hypothetical protein		**Structural head protein**	Not well conserved
13	70	Hypothetical protein		Non-structural protein	No
111	77	Hypothetical protein	nvRNApol β[a]		Yes
26	168	Hypothetical protein		**Structural tail fiber protein**	Yes
110	169	Hypothetical protein		**Structural tail fiber protein**	Yes
24	171	Hypothetical protein		Structural protein	Yes
18 / 19	203	Putative virion protein			Yes
21	214	Hypothetical protein		Structural protein	No
30	219	Hypothetical protein		Non-structural protein	Yes
108	220	Hypothetical protein		Non-structural protein	No
28	238	Hypothetical protein		**Phage infection protein**	Yes
25 / 31 / 109	241	putative DNA-directed RNApol β	vRNApol β[b]		Yes
107	255	Putative virion protein	Tube protein[c]		Yes

[a] Ceyssens et al. 2014
[b] Lee et al. 2011; Ceyssens et al. 2014
[c] Thomas unpublished

7. DISCUSSION

SPN3US and the members of the ϕKZ-related subfamily are a very unusual group of phages which belong to a distant branch of the *Myoviridae* family and they possess a high percentage of functionally unassigned genes. The genetic variation within phages in general is so vast that phages possessing similar characteristics, for example, morphology, replication and general genomic architectures, could be completely unrelated at the nucleotide level (Hendrix et al. 1999). The high divergence of ϕKZ-related phages from other myoviruses and the high amount of phage proteins with unknown functions limits bioinformatics analyses as typically only 20% of the proteins can be assigned a function through sequence comparisons (Thomas et al. 2008) which makes it difficult to find similarities to annotate newly sequenced genomes.

The goals of this project were to identify and assign functions to essential genes in SPN3US using amber mutant phages and to determine if these identified essential genes are well conserved among the ϕKZ-related phages. This will help us create a set of core genes for all ϕKZ-related phages which has not been done before, and by demonstrating this clear relation between SPN3US and the phages of this subfamily we could reliably establish SPN3US as a model whose study will have relevance to all these related complex and diverged bacterial viruses.

Diverse ϕKZ-related phages

The use of phages containing amber mutations has proved to be efficient to identify essential genes in SPN3US. Fifteen genes were analyzed and fourteen were identified as essential genes in this study: gps25, 47, 70, 168, 169, 171, 203, 214, 219, 220, 238, 241 and 255 (Fig 24 shown with rectangles). Previously the general function of three of these genes was known: the genes encoding for gp241which is a vRNAP, gp255 which is a tube protein and gp77 which is a nvRNAP; and two genes were identified as putative virion proteins: gp203 and gp255 (Lee et al. 2011). Now we can infer the function of four more genes: gp47 which is a structural head protein, gp168 and 169 which are baseplate proteins, and gp238 which is a protein associated with host infection (Table 9 and Table 16). We also know that ten of the fifteen analyzed genes are encoding for virion associated proteins, which is a significant proportion. Previously from homology only 27 proteins

were predicted to be structural (Lee et al. 2011), we have now determined that the virion of this phage is complex, encoding about 80 different proteins corresponding to 46% of the genome, and this virion complexity is not only decoration, it is essential for survival.

Ten of these newly identified essential genes are clearly well conserved among the ϕKZ-related phages, the majority encoding virion proteins (Table 16). We suggest they form part of the core genes for this subfamily. A highest similarity was found among the proteins from the proposed SPN3US-like genus (SPN3US, ϕEaH2 and CR5), not only in percentage of identity but also in conservation of gene order (Figs 19, 22 and 23). Within the genus a closer evolutionary relation could be observed with ϕEaH2 (Domotor et al. 2012) as it was the only phage with which similarity was determined at the nucleotide level (Table 8). For the other members of the ϕKZ-related phages similarities were found only at the protein level when using a PSIBLAST search, even for the control proteins whose coding genes are anticipated to be highly conserved among the group. This was expected as there is higher conservation of amino acids than of nucleotides and previous studies had determined that this group has a higher divergence rate than other phages and a very ancient common ancestor (Thomas et al. 2008).

These studies confirm that *Pseudomonas* phage ϕKZ has a similar degree of protein conservation to *Pseudomonas* phage EL and to *Salmonella* phage SPN3US. This is unusual as usually the more closely related phages hosts are, the more closely related the phages are (Hendrix et al. 1999; Grose and Casjens 2014). This diverged nature of the ϕKZ-related phages makes phylogenetic analyses of these phages difficult, ideally as more related phage genomes are sequenced it will be feasible to determine the evolutionary relationships amongst these phages.

Conserved essential genes in the ϕKZ-related phages

The genome architecture of *Salmonella* phage SPN3US shares similarity with other ϕKZ-related genomes as can be observed in Figures 19, 22 and 23. There is a genomic organization in syntenic blocks that contain sets of conserved essential genes, for example in figure 19 we can observe the essential genes coding for subunits of the DNA polymerase, RNA polymerase and nuclease SbcCD in the same region which is maintained in most of the ϕKZ-related phages. We

have determined that the majority of the conserved essential genes identified in SPN3US encode for virion proteins and for some of them we could infer a function.

Gp47 has been identified as a capsid protein due to the mass spectral data which shows evidence of cleavage by the prohead protease (Fig 20). Even though the amino acid sequence of the predicted protein shows no similarity with phages outside the SPN3US-like, as demonstrated in Figure 19, a conservation of gene order is observed. This could suggest that gp47 varies more rapidly.

Gene products gp168 and gp169 have demonstrated a high conservation of the N-terminal domain when aligned with the orthologous proteins in the ϕKZ-related phages (Fig 22). Based on the findings of Sycheva et al. (2012) we can infer that the paralogs present in SPN3US (gp167-170) participate in host cell recognition and are associated with the baseplate tail fibers having the N-terminal domain attached to the phage baseplate while the C-terminal domain, which is not conserved, could bind to diverse receptors in the host cell surface. Outside the subfamily we found weak matches for gp131 in ϕKZ to the putative long tail fiber protein of *Listeria* phage A118 which is a Siphovirus, and more remarkable we have matches to the N-terminal domain of the paralogous proteins gp10 and gp12 in phage T4 which are located in the periphery of the baseplate. As reported by Leiman et al. (2006) the N-terminal domain of the paralogous proteins dock to one another and the C-terminal goes the opposite direction as it will interact with the host cell surface. This statement strengthens our hypothesis that the N-terminal domain of our paralogous proteins are more conserved than the C-terminal domain.

The virion protein gp238 seems to be related with host interaction. Due to the hits found outside the ϕKZ-related subfamily and the prediction of transmembrane protein helices we speculate it is a tail protein, probably in the baseplate as the position of the gene encoding for gp238 is located upstream of the gene encoding for the cell puncturing device protein which can be seen in the genome map in Figure 24.

The essential proteins that are not conserved in all the members of the ϕKZ-related phages might have a host specific function and need further analyses.

Developing a genetic system for SPN3US

We have verified that with the twenty-four-hour hydroxylamine mutagenesis treatment 17 of the 18 amber mutant phage candidates analyzed presented a single amber mutation per genome. Although they may appear in a low rate, phages with two amber mutations could also be expected after the mutagenesis treatment. Once we have confirmed that combining amber mutant phages for sequencing did not always clearly identify all the amber mutations or to assign which phage presented the amber mutation, we need to emphasize the importance of doing individual genome sequencing. Given that the hydroxylamine mutagenesis treatment is a random process individual genome sequencing also allowed us to identify all the other mutations present in each mutant phage. Such information could be critical for downstream analyses.

Another important factor to take into account for future studies will be the rate of repeated amber mutant phages. In the present research four of the previously isolated phages presented an amber mutation in the same location site (amber 25 with 31 and amber 6 with 20) (Table 5.1). When isolating the amber mutant candidate phages two out of the six candidates were duplicates constituting a drawback of the method, thus it is imperative to perform appropriate cross-plating to avoid the characterization of repeated amber mutant phages and ensure that each amber mutant phage is genetically different from the others.

MiSeq is preferred to HiSeq which was previously used, quality of the data is good and economically it is better to go from HiSeq to MiSeq. The high sequence duplication level detected for the HiSeq data suggests that sequencing capacity is being wasted as the same sequences are being resequenced over and over again (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help).

In summary the most appropriate, economical and time efficient approach to identify our essential genes in phage SPN3US is to obtain amber mutant phage candidates with a twenty-four-hour hydroxylamine treatment, cross-plate the isolated amber mutant phage candidates, sequence them individually using a MiSeq system and perform PSIBLAST searches to find the most significant similarities with the ϕKZ-related phages.

An interesting trend observed regarding the isolated amber mutant phages was that the majority of them showed a mutation in the glutamine codon (Table 5.1 and 5.2). Considering that the frequency of occurrence of glutamine and tryptophan codons is very similar (Table 17) a possible explanation of why less tryptophan codons are rescued in comparison with the glutamine codons could be the type of suppressor host and the position of the rescued codon. The suppressor hosts used in this study, UB0017 and TT6675, present a specialized tRNA that inserts a polar amino acid, glutamine and serine respectively, at the amber codon. If the mutated codon is a CAG, the tRNA of the UB0017 host will insert a glutamine codon which will rescue the protein to its original sequence; in the TT6675 host a serine will be inserted instead, and since both are neutral polar amino acids a replacement would be more accepted. On the other hand, tryptophan is unique in terms of chemistry and size so a substitution by either a glutamine or a serine could be deleterious, unless the position were the replacement occurs does not infer with the protein folding. It would be interesting if in the future the efficiency of suppression of the host is analyzed to determine which host will lead to more isolated candidates.

Table 17. Codon table showing the frequency of the different codons in *Salmonella* phage SPN3US. Tryptophan and Glycine codon frequencies can be seen in yellow. Table obtained from the Codon Usage Database, Countcodon program version 4 (http://www.kazusa.or.jp/codon/countcodon.html).

Salmonella phage SPN3US (80137 codons)

fields: [triplet] [frequency: **per thousand**] ([number])

```
UUU 18.7(  1497)  UCU 14.6(  1167)  UAU 15.5(  1244)  UGU 14.9(  1197)
UUC 16.7(  1337)  UCC 11.3(   908)  UAC 16.4(  1315)  UGC 16.1(  1287)
UUA 18.0(  1442)  UCA 13.4(  1072)  UAA 16.0(  1281)  UGA 18.1(  1451)
UUG 16.5(  1324)  UCG 17.2(  1380)  UAG  7.2(   573)  UGG 18.0(  1441)

CUU 14.3(  1148)  CCU 10.5(   838)  CAU 13.4(  1077)  CGU 22.2(  1777)
CUC  9.7(   776)  CCC  8.9(   713)  CAC 15.0(  1199)  CGC 15.2(  1215)
CUA  9.5(   762)  CCA 14.6(  1174)  CAA 17.1(  1371)  CGA 20.0(  1603)
CUG 19.9(  1597)  CCG 15.9(  1272)  CAG 17.4(  1398)  CGG 14.6(  1174)

AUU 16.1(  1288)  ACU 14.3(  1148)  AAU 15.6(  1249)  AGU 13.3(  1068)
AUC 17.8(  1427)  ACC 19.1(  1533)  AAC 22.8(  1824)  AGC 13.9(  1111)
AUA 12.0(   958)  ACA 15.2(  1221)  AAA 21.2(  1699)  AGA 14.1(  1130)
AUG 16.8(  1343)  ACG 22.1(  1770)  AAG 18.3(  1466)  AGG 10.4(   833)

GUU 22.0(  1761)  GCU 13.8(  1107)  GAU 17.9(  1437)  GGU 18.8(  1508)
GUC 13.9(  1110)  GCC 11.5(   924)  GAC 15.6(  1250)  GGC 13.4(  1072)
GUA 16.2(  1299)  GCA 15.8(  1270)  GAA 23.8(  1906)  GGA 14.0(  1118)
GUG 16.2(  1295)  GCG 16.7(  1339)  GAG 10.1(   810)  GGG 10.6(   853)
```

7.1 FUTURE WORK

The specific functions of the essential genes identified in this study still needs to be determined, thus more experiments including proteomics and transcriptomics approaches should be performed.

In order to identify and determine the function of all essential genes in *Salmonella* phage SPN3US, more amber mutant phages need to be isolated and sequenced. A comparison of different mutagens could be performed to test the effectiveness in the production of amber mutants in SPN3US. Previous studies in T4 bacteriophage have used the mutagens 2-amino purine (2-AP), ethyl methanesulfonate (EMS), 5-bromodeoxyuridine (5-BdU), nitrous acid and proflavine (Stahl 1995).

8. CONCLUSIONS

General mutagenesis with hydroxylamine has proven to be efficient to obtain a single amber mutation per genome in *Salmonella* phage SPN3US in order to identify essential genes. From the fourteen essential genes identified in this study ten were found to be well conserved among the ϕKZ-related phages, and thus they could be classified as core genes of this subfamily. Structural functions were found for the majority of the genes and four genes have now suggested functions.

This research is an important contribution to the understanding of these giant phages, not only for the newly assigned functions but also for the development of an efficient process to analyze other phages to identify the functions of their essential genes. Studies on SPN3US will be relevant to the entire subfamily of phages as it could be considered the first genetic model for giant phages

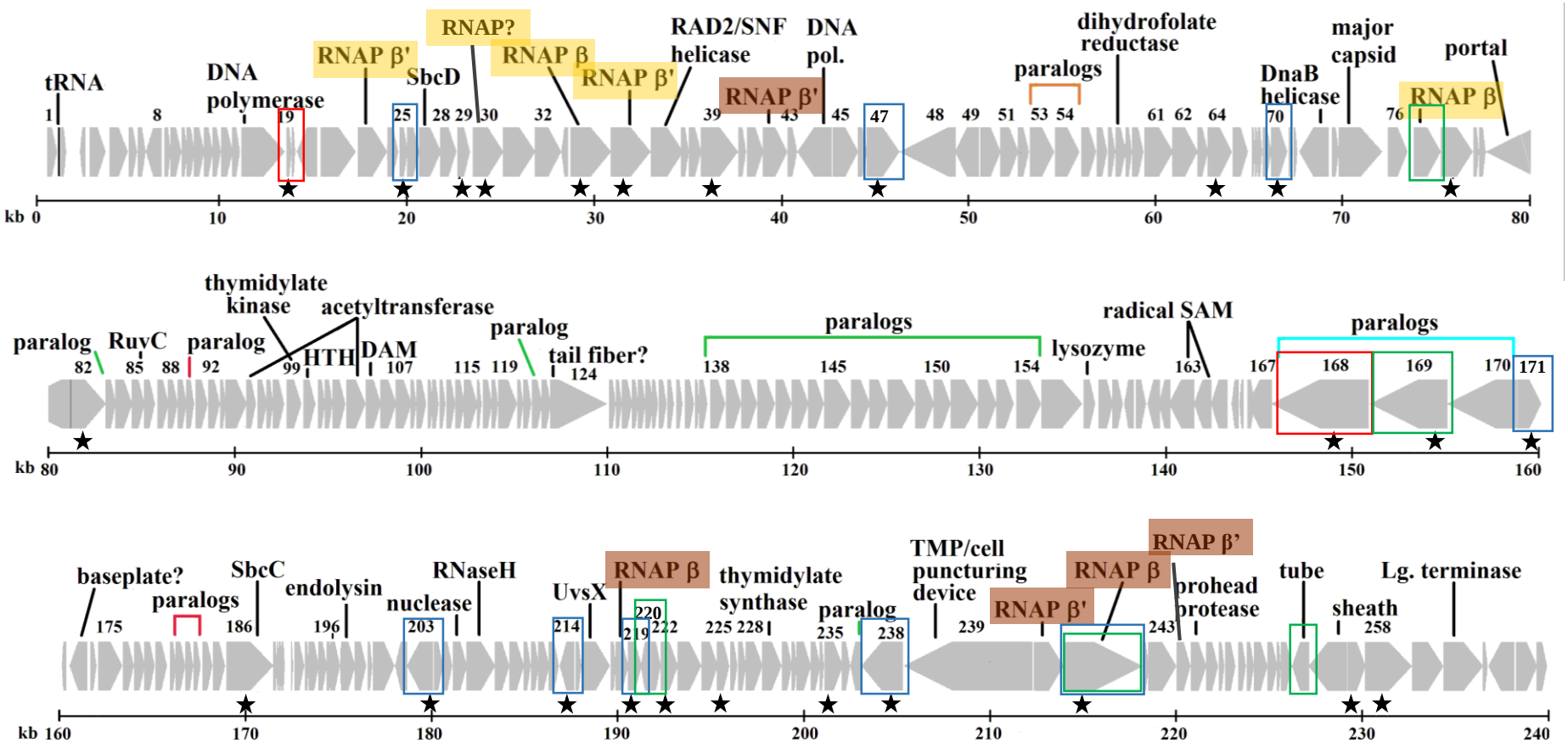

Figure 24: *Salmonella* phage SPN3US genome map showing the genes presenting an amber mutation site. Blue and red rectangles indicate the samples that underwent confirmatory Sanger sequencing, blue are amber mutants that presented a single mutation and red indicate the double mutation present in a single amber mutant phage. Green rectangles indicate the proteins corresponding to the amber mutant phages that were sequenced by MiSeq system and only PSIBLAST was employed to analyze homology. RNAP β and β'-like subunits colored in brown are part of the virion and the subunits colored in yellow are not virion-associated proteins. ★ denote proteins in which an amber mutation was identified from a mix of 50 amber mutant candidates previously sequenced.

9. LITERATURE CITED

Ackermann HW. 1998. Tailed bacteriophages: the order *Caudovirales*. Adv. Virus Res. 51:135–201.

Ackermann H-W, Prangishvili D. 2012. Prokaryote viruses studied by electron microscopy. Arch. Virol. 157:1843–9.

Allen HK, Trachsel J, Looft T, Casey T. 2014. Finding alternatives to antibiotics. Ann. N. Y. Acad. Sci.:1–10.

Anderson B, Rashid MH, Carter C, Pasternack G, Rajanna C, Revazishvili T, Dean T, Senecal A, Sulakvelidze A. 2011. Enumeration of bacteriophage particles: Comparative analysis of the traditional plaque assay and real-time QPCR- and NanoSight-based assays. Bacteriophage 1:86–93.

Black LW, Thomas JA. 2012. Condensed genome structure. Rossmann MG, Rao VB, editors. Adv Exp Med Biol 726:469–487.

Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F. 2002. Genomic analysis of uncultured marine viral communities. Proc. Natl. Acad. Sci. U. S. A. 99:14250–14255.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST plus : architecture and applications. BMC Bioinformatics 10.

Ceyssens P-J, Minakhin L, Van den Bossche A, Yakunina M, Klimuk E, Blasdel B, De Smet J, Noben J-P, Bläsi U, Severinov K, et al. 2014. Development of giant bacteriophage ϕKZ is independent of the host transcription apparatus. J. Virol. 88:10501–10.

Chibani-Chennoufi S, Bruttin A, Dillmann M, Bru H. 2004. Phage-Host interaction : an ecological perspective. J. Bacteriol. 186:3677–3686.

Cock PJ a, Antao T, Chang JT, Chapman B a, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422–3.

Comeau AM, Bertrand C, Letarov A, Tétart F, Krisch HM. 2007. Modular architecture of the T4 phage superfamily: A conserved core genome and a plastic periphery. Virology 362:384–396.

Comeau AM, Hatfull GF, Krisch HM, Lindell D, Mann NH, Prangishvili D. 2008. Exploring the prokaryotic virosphere. Res. Microbiol. 159:306–313.

Cornelissen A., Hardies SC, Shaburova O V., Krylov VN, Mattheus W, Kropinski A. M, Lavigne

R. 2012. Complete genome sequence of the giant virus OBP and comparative genome analysis of the diverse ϕKZ-related phages. J. Virol. 86:1844–1852.

Domotor D, Becsagh P, Rakhely G, Schneider G, Kovacs T. 2012. Complete genomic sequence of *Erwinia amylovora* phage PhiEaH2. J. Virol. 86:10899–10899.

Epstein RH, Bolle A, Steinberg CM. 2012. Amber mutants of bacteriophage T4D: Their isolation and genetic characterization. Genetics 190:831–840.

Fresse E, Bautz E, Bautz E. 1961. The chemical and mutagenic specificity of hydroxylamine. Proc. Natl. Acad. Sci. U. S. A. 47:845–855.

Frieden T. 2013. Antibiotic resistance threats. Cdc:22–50.

Grassberger, Martin, Ronald A. Sherman, Olga S. Gileva, Christopher MH Kim and KYM. 2013. Biotherapy-history, principles and practice: a practical guide to the diagnosis and treatment of disease using living organisms. Springer Science & Business Media.

Grose JH, Casjens SR. 2014. Understanding the enormous diversity of bacteriophages: The tailed phages that infect the bacterial family Enterobacteriaceae. Virology 468-470C:421–443.

Haq IU, Chaudhry WN, Akhtar MN, Andleeb S, Qadri I. 2012. Bacteriophages and their implications on future biotechnology: A Review. Virol. J. 9:9.

Harper DR, Enright MC. 2011. Bacteriophages for the treatment of *Pseudomonas aeruginosa* infections. J. Appl. Microbiol. 111:1–7.

Hatfull GF. 2008. Bacteriophage Genomics. Curr Opin Microbiol. 11:447–453.

Hatfull GF, Hendrix RW. 2011. Bacteriophages and their genomes. Curr. Opin. Virol. 1:298–303.

Hendrix RW, Smith MC, Burns RN, Ford ME, Hatfull GF. 1999. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. Proc. Natl. Acad. Sci. U. S. A. 96:2192–2197.

Jones DT. 1999. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292:195–202.

Jover LF, Effler TC, Buchan A, Wilhelm SW, Weitz JS. 2014. The elemental composition of virus particles: implications for marine biogeochemical cycles. Nat. Rev. Microbiol. 12:519–28.

Kropinski AM, Mazzocco A, Waddell TE, Lingohr E, Johnson RP. 2009. Enumeration of bacteriophages by double agar overlay plaque assay. Methods Mol. Biol. 501:69–76.

Krupovic M, Prangishvili D, Hendrix RW, Bamford DH. 2011. Genomics of bacterial and archaeal viruses: Dynamics within the prokaryotic virosphere. Microbiol. Mol. Biol. Rev. 75:610–635.

Krylov V, Shaburova O, Krylov S, Pleteneva E. 2013. A genetic approach to the development of new therapeutic phages to fight *Pseudomonas aeruginosa* in wound infections. Viruses 5:15–53.

Krylov VN, Dela Cruz DM, Hertveldt K, Ackermann H-W. 2007. "ϕKZ-like viruses", a proposed new genus of myovirus bacteriophages. Arch. Virol. 152:1955–9.

Krylov VN, Smirnova TA, Minenkova IB, Plotnikova TG, Zhazikov IZ, Khrenova EA. 1984. *Pseudomonas* bacteriophage ϕKZ contains an inner body in its capsid. Can. J. Microbiol. 30:758–762.

Lavigne R, Darius P, Summer EJ, Seto D, Mahadevan P, Nilsson AS, Ackermann HW, Kropinski AM. 2009. Classification of *Myoviridae* bacteriophages using protein sequence similarity. BMC Microbiol. 9:224.

Lee J-H, Shin H, Kim H, Ryu S. 2011. Complete genome sequence of *Salmonella* bacteriophage SPN3US. J. Virol. 85:13470–1.

Leiman PG, Shneider MM, Mesyanzhinov V V., Rossmann MG. 2006. Evolution of bacteriophage tails: Structure of T4 gene product 10. J. Mol. Biol. 358:912–921.

Mann NH. 2005. The third age of phage. PLoS Biol. 3:e182.

Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z, et al. 2005. CDD: A Conserved Domain Database for protein classification. Nucleic Acids Res. 33:192–196.

Mead PS, Slutsker L, Dietz V, McCaig LF, Bresee JS, Shapiro C, Griffin PM, Tauxe R V. 1999. Food-related illness and death in the United States. Emerg. Infect. Dis. 5:607–625.

Mesyanzhinov V V, Robben J, Grymonprez B, Kostyuchenko V a, Bourkaltseva M V, Sykilinda NN, Krylov VN, Volckaert G. 2002. The genome of bacteriophage ϕKZ of *Pseudomonas aeruginosa*. J. Mol. Biol. 317:1–19.

Nelson EJ, Harris JB, Morris JG, Calderwood SB, Camilli A. 2009. Cholera transmission: the host, pathogen and bacteriophage dynamic. Nat. Rev. Microbiol. 7:693–702.

Parracho HM, Burrowes BH, Enright MC, McConville ML, Harper DR. 2012. The role of regulated clinical trials in the development of bacteriophage therapeutics. J. Mol. Genet. Med. 6:279–86.

Sano E, Carlson S, Wegley L, Rohwer F. 2004. Movement of viruses between biomes. Appl. Environ. Microbiol. 70:5842–6.

Serwer P, Hayes SJ, Thomas J a, Hardies SC. 2007. Propagating the missing bacteriophages: a large bacteriophage in a new class. Virol. J. 4:21.

Serwer P, Hayes SJ, Zaman S, Lieman K, Rolando M, Hardies SC. 2004. Improved isolation of undersampled bacteriophages: Finding of distant terminase genes. Virology 329:412–424.

Shin H, Lee J-H, Kim H, Choi Y, Heu S, Ryu S. 2012. Receptor diversity and host interaction of bacteriophages infecting *Salmonella enterica* serovar Typhimurium. PLoS One 7:e43392.

Söding J. 2005. Protein homology detection by HMM-HMM comparison. Bioinformatics 21:951–960.

Sokolova OS, Shaburova O V, Pechnikova E V, Shaytan AK, Krylov S V, Kiselev NA, Krylov VN. 2014. Genome packaging in EL and Lin68, two giant ϕKZ-like bacteriophages of *P. aeruginosa*. Virology 468-470C:472–478.

Sonnhammer EL, von Heijne G, Krogh a. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc. Int. Conf. Intell. Syst. Mol. Biol. 6:175–182.

Stahl FW. 1995. The amber mutants of phage T4. Genetics 141:439.

Sulakvelidze A, Alavidze Z, Morris JG. 2001. Bacteriophage therapy. Antimicrob. Agents Chemother. 45:649–659.

Sycheva L V, Shneider MM, Sykilinda NN, Ivanova MA, Miroshnikov KA, Leiman PG. 2012. Crystal structure and location of gp131 in the bacteriophage ϕKZ virion. Virology 434:257–64.

Tessman I. 1968. Mutagenic treatment of double- and single-stranded DNA phages T4 and S13 with hydroxylamine. Virology 35:330–333.

Thomas J a, Weintraub ST, Hakala K, Serwer P, Hardies SC. 2010. Proteome of the large *Pseudomonas* myovirus 201phi2-1: delineation of proteolytically processed virion proteins. Mol. Cell. Proteomics 9:940–951.

Thomas JA. 2015. Phage Biology Laboratory Manual. RIT course Bio 335.

Thomas JA, Black LW. 2013. Mutational analysis of the *Pseudomonas aeruginosa* myovirus ϕKZ morphogenetic protease gp175. J. Virol. 87:8713–25.

Thomas JA, Rolando MR, Carroll CA, Shen PS, Belnap DM, Weintraub ST, Serwer P, Hardies SC. 2008. Characterization of *Pseudomonas chlororaphis* myovirus 201ϕ2-1 via genomic sequencing, mass spectrometry, and electron microscopy. Virology 376:330–338.

Thomas JA, Weintraub ST, Wu W, Winkler DC, Cheng N, Steven AC, Black LW. 2012. Extensive proteolysis of head and inner body proteins by a morphogenetic protease in the giant *Pseudomonas aeruginosa* phage ϕKZ. Mol. Microbiol. 84:324–39.

Verbeken G, De Vos D, Vaneechoutte M, Merabishvili M, Zizi M, Pirnay J-P. 2007. European regulatory conundrum of phage therapy. Future Microbiol. 2:485–491.

(n.d.). Retrieved February 20, 2015, from
http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=28883

10 APPENDIX

Appendix A. SNP summary reports for amber mutant candidates sequenced by MiSeq system.

Table A.1. SNP summary report for amber mutant candidate 107. Nonsense mutation is highlighted in yellow.

Contig ID	Ref Pos	Type	Ref Base	Called Base	Impact	SNP %	P not ref	Feature Type	Feature Name	DNA Change	Amino Acid Change	Depth
JN641803	27612	SNP	C	T	Non Synonymous	99.60%	100.00%	CDS	SPN3US_0032	c.394C>T	p.P132S	884
JN641803	53716	SNP	G	A	Synonymous	100.00%	100.00%	CDS	SPN3US_0053	c.111G>A	p.Q37Q	922
JN641803	61708	SNP	G	A	Synonymous	100.00%	100.00%	CDS	SPN3US_0062	c.519G>A	p.L173L	876
JN641803	63491	SNP	G	A	Synonymous	99.70%	100.00%	CDS	SPN3US_0064	c.405G>A	p.E135E	791
JN641803	69613	SNP	G	A	Non Synonymous	99.70%	100.00%	CDS	SPN3US_0074	c.46G>A	p.D16N	790
JN641803	82224	SNP	C	T	Non Synonymous	100.00%	100.00%	CDS	SPN3US_0082	c.1498C>T	p.R500C	845
JN641803	109100	SNP	C	T	Non Synonymous	99.80%	100.00%	CDS	SPN3US_0124	c.2275C>T	p.L759F	837
JN641803	119591	SNP	G	A	Synonymous	99.40%	100.00%	CDS	SPN3US_0142	c.690G>A	p.P230P	584
JN641803	126818	SNP	G	A	Synonymous	100.00%	100.00%	CDS	SPN3US_0149	c.423G>A	p.L141L	727
JN641803	130321	SNP	G	A	Synonymous	99.20%	100.00%	CDS	SPN3US_0152	c.570G>A	p.G190G	772
JN641803	136018	SNP	G	A	Synonymous	99.70%	100.00%	CDS	SPN3US_0156	c.594G>A	p.V198V	794
JN641803	153287	SNP	C	T	Synonymous	99.50%	100.00%	CDS	SPN3US_0169	c.1659G>A	p.A553A	716
JN641803	185132	SNP	G	A	Non Synonymous	100.00%	100.00%	CDS	SPN3US_0210	c.292G>A	p.G98S	1000
JN641803	200944	Ins	-	T		96.10%	100.00%					723
JN641803	223871	SNP	C	A	Non Synonymous	99.60%	100.00%	CDS	SPN3US_0250	c.70C>A	p.R24S	899
JN641803	226689	SNP	G	A	Nonsense	99.80%	100.00%	CDS	SPN3US_0255	c.448C>T	p.Q150.	906
JN641803	231556	SNP	G	A	Synonymous	100.00%	100.00%	CDS	SPN3US_0258	c.1398G>A	p.V466V	910
JN641803	239933	Ins	-	T		99.50%	100.00%					240

Table A.2. SNP summary report for amber mutant candidate 108. Nonsense mutation is highlighted in yellow

Contig ID	Ref Pos	Type	Ref Base	Called Base	Impact	SNP %	P not ref	Feature Type	Feature Name	DNA Change	Amino Acid Change	Depth
JN641803	16926	SNP	G	A	Non Synonymous	100.00%	100.00%	CDS	SPN3US_0022	c.1159G>A	p.V387I	491
JN641803	32268	SNP	G	A	Non Synonymous	99.70%	100.00%	CDS	SPN3US_0035	c.1018G>A	p.A340T	448
JN641803	77962	Ins	-	T		56.20%	90.00%					338
JN641803	87718	Ins	-	C		74.00%	100.00%					277
JN641803	109947	SNP	G	A		100.00%	100.00%					442
JN641803	169781	SNP	G	A	Synonymous	99.70%	100.00%	CDS	SPN3US_0186	c.882G>A	p.S294S	462
JN641803	191499	SNP	G	A	Nonsense	100.00%	100.00%	CDS	SPN3US_0220	c.76C>T	p.Q26.	504
JN641803	200944	Ins	-	T		97.20%	100.00%					398
JN641803	223871	SNP	C	A	Non Synonymous	100.00%	100.00%	CDS	SPN3US_0250	c.70C>A	p.R24S	518
JN641803	239933	Ins	-	T		100.00%	100.00%					220

Table A.3. SNP summary report for amber mutant candidate 109. Nonsense mutation is highlighted in yellow

Contig ID	Ref Pos	Type	Ref Base	Called Base	Impact	SNP %	P not ref	Feature Type	Feature Name	DNA Change	Amino Acid Change	Depth
JN641803	77962	Ins	-	T		95.30%	100.00%					365
JN641803	87722	Ins	-	C		70.60%	100.00%					307
JN641803	93451	Ins	-	T		96.80%	100.00%					473
JN641803	123966	SNP	G	A	Non Synonymous	99.80%	100.00%	CDS	SPN3US_0147	c.17G>A	p.S6N	513
JN641803	125357	SNP	G	A	Synonymous	99.30%	100.00%	CDS	SPN3US_0148	c.471G>A	p.P157P	583
JN641803	200944	Ins[2]	-	TT		87.30%	100.00%			g.200944_200945insTT		446
JN641803	211353	SNP	G	A	Synonymous	100.00%	100.00%	CDS	SPN3US_0239	c.975C>T	p.G325G	564
JN641803	214121	SNP	C	T	Nonsense	100.00%	100.00%	CDS	SPN3US_0241	c.133C>T	p.Q45.	443
JN641803	223851	SNP	T	C	Non Synonymous	99.80%	100.00%	CDS	SPN3US_0250	c.50T>C	p.L17P	591
JN641803	239933	Ins	-	T		99.30%	100.00%					144

Table A.4. SNP summary report for amber mutant candidate 110. Nonsense mutation is highlighted in yellow

Contig ID	Ref Pos	Type	Ref Base	Called Base	Impact	SNP %	P not ref	Feature Type	Feature Name	DNA Change	Amino Acid Change	Depth
JN641803	4027	SNP	G	A	Non Synonymous	99.80%	100.00%	CDS	SPN3US_0004	c.571G>A	p.V191I	689
JN641803	77962	Ins	-	T		61.30%	90.00%					326
JN641803	87721	Ins	-	C		79.80%	100.00%					497
JN641803	93451	Ins	-	T		94.40%	100.00%					508
JN641803	123194	SNP	G	A	Non Synonymous	99.70%	100.00%	CDS	SPN3US_0146	c.253G>A	p.G85S	747
JN641803	133130	SNP	G	T	Synonymous	100.00%	100.00%	CDS	SPN3US_0154	c.1296G>T	p.P432P	15
JN641803	133131	SNP	G	A	Non Synonymous	100.00%	100.00%	CDS	SPN3US_0154	c.1297G>A	p.V433I	15
JN641803	133132	SNP	T	A	Non Synonymous	100.00%	100.00%	CDS	SPN3US_0154	c.1298T>A	p.V433D	14
JN641803	133133	SNP	C	A	Synonymous	100.00%	100.00%	CDS	SPN3US_0154	c.1299C>A	p.V433V	9
JN641803	135302	SNP	T	A	Non Synonymous	98.00%	100.00%	CDS	SPN3US_0155	c.2134T>A	p.S712T	516
JN641803	152931	SNP	C	T	Nonsense	99.60%	100.00%	CDS	SPN3US_0169	c.2015G>A	p.W672.	554
JN641803	202483	SNP	C	T	Non Synonymous	99.60%	100.00%	CDS	SPN3US_0237	c.433G>A	p.G145S	528
JN641803	216546	SNP	A	G	Non Synonymous	99.80%	100.00%	CDS	SPN3US_0241	c.2558A>G	p.D853G	763
JN641803	223871	SNP	C	A	Non Synonymous	99.80%	100.00%	CDS	SPN3US_0250	c.70C>A	p.R24S	757
JN641803	234437	SNP	C	T	Synonymous	100.00%	100.00%	CDS	SPN3US_0260	c.75C>T	p.F25F	687
JN641803	239933	Ins	-	T		100.00%	100.00%					103

Table A.5. SNP summary report for amber mutant candidate 111. Nonsense mutation is highlighted in yellow

Contig ID	Ref Pos	Type	Ref Base	Called Base	Impact	SNP %	P not ref	Feature Type	Feature Name	DNA Change	Amino Acid Change	Depth
JN641803	67	SNP	C	T		100.00%	100.00%					60
JN641803	62383	SNP	G	A	Synonymous	99.80%	100.00%	CDS	SPN3US_0062	c.1194G>A	p.S398S	853
JN641803	74357	SNP	C	T	Nonsense	99.80%	100.00%	CDS	SPN3US_0077	c.604C>T	p.Q202.	813
JN641803	223871	SNP	C	A	Non Synonymous	99.80%	100.00%	CDS	SPN3US_0250	c.70C>A	p.R24S	816
JN641803	239933	Ins	-	T		98.10%	100.00%					158

Appendix B. Summary of BLAST analyses for identified essential genes in SPN3US and their products among the ϕKZ-related phages using different BLAST tools.

Table B.1. Hits for gp70 in SPN3US at the nucleotide and protein level using different BLAST tools. - refers to no match identified.

Phage	BLASTn	tBLASTn	BLASTp	PSIBLAST
Erwinia phage ϕEaH2	-	✓	✓	✓
Cronobacter phage CR5	-	-	-	-
Pseudomonas phage ϕKZ	-	-	-	-
Pseudomonas phage 201ϕ2-1	-	-	-	-
Pseudomonas phage ϕPA3	-	-	-	-
Erwinia phage Ea35-70	-	-	-	-
Erwinia phage ϕEaH1	-	-	-	-
Ralstonia phage RSL2	-	-	-	-
Vibrio phage JM-2012	-	-	-	-
Vibrio phage VP4B	-	-	-	-
Pseudomonas phage OBP	-	-	-	-
Pseudomonas phage EL	-	-	-	-

Table B.2. Hits for gp168 in SPN3US at the nucleotide and protein level using different BLAST tools. - refers to no match identified

Phage	BLASTn	tBLASTn	BLASTp	PSIBLAST
Erwinia phage ϕEaH2	✓	✓	✓	✓
Cronobacter phage CR5	-	✓	✓	✓
Pseudomonas phage ϕKZ	-	✓	✓	✓
Pseudomonas phage 201ϕ2-1	-	✓	✓	✓
Pseudomonas phage ϕPA3	-	✓	✓	✓
Erwinia phage Ea35-70	-	✓	✓	✓
Erwinia phage ϕEaH1	-	✓	✓	✓
Ralstonia phage RSL2	-	✓	✓	✓
Vibrio phage JM-2012	-	-	✓	✓
Vibrio phage VP4B	-	-	✓	✓
Pseudomonas phage OBP	-	✓	✓	✓
Pseudomonas phage EL	-	✓	✓	✓

Table B.3. Hits for gp171 in SPN3US at the nucleotide and protein level using different BLAST tools. - refers to no match identified

Phage	BLASTn	tBLASTn	BLASTp	PSIBLAST
Erwinia phage ϕEaH2	✓	✓	✓	✓
Cronobacter phage CR5	-	✓	✓	✓
Pseudomonas phage ϕKZ	-	✓	✓	✓
Pseudomonas phage 201ϕ2-1	-	✓	✓	✓
Pseudomonas phage ϕPA3	-	✓	✓	✓
Erwinia phage Ea35-70	-	✓	✓	✓
Erwinia phage ϕEaH1	-	-	✓	✓
Ralstonia phage RSL2	-	-	✓	✓
Vibrio phage JM-2012	-	-	-	-
Vibrio phage VP4B	-	✓	✓	✓
Pseudomonas phage OBP	-	✓	✓	✓
Pseudomonas phage EL	-	✓	✓	✓

Table B.4. Hits for gp203 in SPN3US at the nucleotide and protein level using different BLAST tools. - refers to no match identified

Phage	BLASTn	tBLASTn	BLASTp	PSIBLAST
Erwinia phage ϕEaH2	✓	✓	✓	✓
Cronobacter phage CR5	-	✓	✓	✓
Pseudomonas phage ϕKZ	-	✓	✓	✓
Pseudomonas phage 201ϕ2-1	-	✓	✓	✓
Pseudomonas phage ϕPA3	-	✓	✓	✓
Erwinia phage Ea35-70	-	✓	✓	✓
Erwinia phage ϕEaH1	-	✓	✓	✓
Ralstonia phage RSL2	-	✓	✓	✓
Vibrio phage JM-2012	-	✓	✓	✓
Vibrio phage VP4B	-	✓	✓	✓
Pseudomonas phage OBP	-	✓	✓	✓
Pseudomonas phage EL	-	✓	✓	✓

Table B.5. Hits for gp214 in SPN3US at the nucleotide and protein level using different BLAST tools. - refers to no match identified

Phage	BLASTn	tBLASTn	BLASTp	PSIBLAST
Erwinia phage ϕEaH2	-	✓	✓	✓
Cronobacter phage CR5	-	✓	✓	✓
Pseudomonas phage ϕKZ	-	✓*	✓	✓
Pseudomonas phage 201ϕ2-1	-	-	✓	✓
Pseudomonas phage ϕPA3	-	-	✓	✓
Erwinia phage Ea35-70	-	-	-	-
Erwinia phage ϕEaH1	-	✓*	✓	✓
Ralstonia phage RSL2	-	-	-	✓
Vibrio phage JM-2012	-	-	-	✓
Vibrio phage VP4B	-	-	-	-
Pseudomonas phage OBP	-	-	-	-
Pseudomonas phage EL	-	-	-	-

* Defined by synteny analysis in hits with higher e-values

Table B.6. Hits for gp219 in SPN3US at the nucleotide and protein level using different BLAST tools. - refers to no match identified

Phage	BLASTn	tBLASTn	BLASTp	PSIBLAST
Erwinia phage ϕEaH2	-	✓	✓	✓
Cronobacter phage CR5	-	✓	✓	✓
Pseudomonas phage ϕKZ	-	✓	✓	✓
Pseudomonas phage 201ϕ2-1	-	✓	✓	✓
Pseudomonas phage ϕPA3	-	✓	✓	✓
Erwinia phage Ea35-70	-	✓	✓	✓
Erwinia phage ϕEaH1	-	✓	✓	✓
Ralstonia phage RSL2	-	-	✓	✓
Vibrio phage JM-2012	-	✓	✓	✓
Vibrio phage VP4B	-	✓*	✓	✓
Pseudomonas phage OBP	-	-	-	✓
Pseudomonas phage EL	-	-	-	✓

* Defined by synteny analysis in hits with higher e-values

Table B.7. Hits for gp238 in SPN3US at the nucleotide and protein level using different BLAST tools. - refers to no match identified

Phage	**BLASTn**	**tBLASTn**	**BLASTp**	**PSIBLAST**
Erwinia phage ϕEaH2	✓	✓	✓	✓
Cronobacter phage CR5	-	✓	✓	✓
Pseudomonas phage ϕKZ	-	✓	✓	✓
Pseudomonas phage 201ϕ2-1	-	✓	✓	✓
Pseudomonas phage ϕPA3	-	✓	✓	✓
Erwinia phage Ea35-70	-	✓	✓	✓
Erwinia phage ϕEaH1	-	✓	✓	✓
Ralstonia phage RSL2	-	✓	✓	✓
Vibrio phage JM-2012	-	-	✓	✓
Vibrio phage VP4B	-	✓	✓	✓
Pseudomonas phage OBP	-	✓	✓	✓
Pseudomonas phage EL	-	✓	✓	✓

Table B.8. Hits for gp241 in SPN3US at the nucleotide and protein level using different BLAST tools. - refers to no match identified

Phage	**BLASTn**	**tBLASTn**	**BLASTp**	**PSIBLAST**
Erwinia phage ϕEaH2	-	✓	✓	✓
Cronobacter phage CR5	-	✓	✓	✓
Pseudomonas phage ϕKZ	-	✓	✓	✓
Pseudomonas phage 201ϕ2-1	-	✓	✓	✓
Pseudomonas phage ϕPA3	-	✓	✓	✓
Erwinia phage Ea35-70	-	✓	✓	✓
Erwinia phage ϕEaH1	-	✓	✓	✓
Ralstonia phage RSL2	-	✓	✓	✓
Vibrio phage JM-2012	-	✓	✓	✓
Vibrio phage VP4B	-	✓	✓	✓
Pseudomonas phage OBP	-	✓	✓	✓
Pseudomonas phage EL	-	✓	✓	✓

Appendix C. Hits found for the identified essential proteins in SPN3US among the ϕKZ-related phages using PSIBLAST

Table C.1. PSIBLAST hits for similar proteins of SPN3US gp25 in the ϕKZ-related phages

Phage	Accession number	Title	aa length	% id	Query coverage	E-value	Bitscore
Erwinia phage ϕEaH2	YP_007237857.1	hypothetical protein	128	69.05	98	3.00E-38	136
Cronobacter phage CR5	YP_008125766.1	hypothetical protein CR5_026	124	51.24	95	2.00E-33	124
Pseudomonas phage ϕKZ	NP_803628.1	ORF062	136	20	98	1.00E-32	122
Pseudomonas phage 201ϕ2-1	YP_001956842.1	hypothetical protein 201phi2-1p118	89	21.43	66	0.002	43.6
Pseudomonas phage ϕPA3	AEH03483.1	hypothetical protein	89	15.29	66	0.54	37.1
Erwinia phage Ea35-70	YP_009005029.1	hypothetical protein Ea357_234	134	24.22	98	3.00E-36	132
Erwinia phage ϕEaH1	YP_009010072.1	hypothetical protein	142	16.67	98	1.00E-33	125
Ralstonia phage RSL2	BAQ02747.1	hypothetical protein	138	17.69	100	4.00E-30	116
Vibrio phage JM-2012	YP_006383418.1	hypothetical protein TSMG0138	130	16.95	92	2.00E-05	49.8
Vibrio phage VP4B	AGB07197.1	hypothetical protein	141	19.69	98	2.00E-31	119
*Pseudomona*s phage OBP	YP_004957970.1	unnamed protein product	136	21.71	98	5.00E-29	113
Pseudomonas phage EL	YP_418061.1	hypothetical protein PPEV_gp028	135	25.6	96	4.00E-09	60.5

Table C.2. PSIBLAST hits for similar proteins of SPN3US gp47 in the ϕKZ-related phages

Phage	Accession number	Title	aa length	% id	Query coverage	E-value	Bitscore
Erwinia phage ϕEaH2	YP_007237881.1	hypothetical protein	566	56.34	99	0	787
Cronobacter phage CR5	YP_008125788.1	hypothetical protein CR5_048	566	26.82	98	0	693

Table C.3. BLASTp hits for SPN3US gp47 with e-values higher than the stringent threshold and located in a synteny region shared in the ϕKZ-related phages

Phage	Accession number	Title	aa length	% id	E-value	Bitscore
Erwinia phage Ea35-70	YP_009005081.1	hypothetical protein Ea357_285	418	25	0.008	32
Erwinia phage ϕEaH1	YP_009010285.1	hypothetical protein CF95_gp232	456	26.09	0.013	30.8

Table C.4. PSIBLAST hits for similar proteins of SPN3US gp70 in the ϕKZ-related phages

Phage	Accession number	Title	aa length	% id	Query coverage	E-value	Bitscore
Erwinia phage ϕEaH2	YP_007237907.1	hypothetical protein	292	29.24	80	5.00E-119	355

Table C.5. PSIBLAST hits for similar proteins of SPN3US gp77 in the ϕKZ-related phages

Phage	Accession number	Protein title	aa length	% id.	Query Coverage	E-value	Bitscore
Erwinia phage ϕEaH2	YP_007237652.1	hypothetical protein	591	82.26	100	0	684
Cronobacter phage CR5	YP_008125813.1	hypothetical protein CR5_073	524	57.5	88	0	551
Pseudomonas phage ϕKZ	NP_803689.1	ORF123	543	20.71	99	5.00E-177	514
Pseudomonas phage 201ϕ2-1	YP_001956926.1	hypothetical protein 201phi2-1p203	544	17.89	99	4.00E-176	511
Pseudomonas phage ϕPA3	AEH03562.1	hypothetical protein	540	19.66	99	0	531
Erwinia phage Ea35-70	YP_009004794.1	hypothetical protein Ea357_002	562	16.89	99	2.00E-154	456
Erwinia phage ϕEaH1	YP_009010266.1	hypothetical protein	554	20.2	98	6.00E-157	462
Ralstonia phage RSL2	BAQ02643.1	hypothetical protein	559	19.17	98	7.00E-158	465
Vibrio phage JM-2012	AFI55320.1	hypothetical protein TSMG0037	543	19.9	97	7.00E-143	426
Vibrio phage VP4B	AGB07259.1	hypothetical protein	559	15.81	99	6.00E-149	442
*Pseudomona*s phage OBP	YP_004958033.1	unnamed protein product	563	15.95	99	7.00E-138	414
Pseudomonas phage EL	YP_418113.1	hypothetical protein PPEV_gp080	570	15.22	98	2.00E-149	444

Table C.6. PSIBLAST hits for similar proteins of SPN3US gp171 in the ϕKZ-related phages

Phage	Accession number	Title	aa length	% id	Query coverage	E-value	Bitscore
Erwinia phage ϕEaH2	YP_007237743.1	hypothetical protein	420	80.95	100	6.00E-173	503
Cronobacter phage CR5	YP_008125902.1	hypothetical protein CR5_162	415	46.43	99	2.00E-152	451
Pseudomonas phage ϕKZ	NP_803696.1	ORF130	427	19.76	96	2.00E-136	410
Pseudomonas phage 201ϕ2-1	YP_001956938.1	virion structural protein	428	22.6	93	8.00E-134	403
Pseudomonas phage ϕPA3	AEH03572.1	virion structural protein	426	20.23	93	3.00E-134	404
Erwinia phage Ea35-70	YP_009004808.1	hypothetical protein Ea357_016	409	21.72	92	5.00E-115	354
Erwinia phage ϕEaH1	YP_009010247.1	virion structural protein	417	19.53	88	3.00E-94	301
Ralstonia phage RSL2	BAQ02591.1	hypothetical protein	425	16.42	94	3.00E-108	337
Vibrio phage JM-2012							
Vibrio phage VP4B	AGB07275.1	hypothetical protein	397	19.65	92	2.00E-81	268
*Pseudomona*s phage OBP	YP_004958048.1	unnamed protein product	408	18.93	96	4.00E-66	228
Pseudomonas phage EL	YP_418145.1	hypothetical protein PPEV_gp112	413	17.63	92	1.00E-76	256

Table C.7. PSIBLAST hits for similar proteins of SPN3US gp203 in the ϕKZ-related phages

Phage	Accession number	Title	aa length	% id	Query coverage	E-value	Bitscore
Erwinia phage ϕEaH2	YP_007237774.1	putative virion structural protein	455	80.49	100	0	540
Cronobacter phage CR5	YP_008125922.1	putative virion structural protein 16	445	56.86	98	2.00E-169	496
Pseudomonas phage ϕKZ	NP_803723.1	ORF157	445	29.78	98	4.00E-157	465
Pseudomonas phage 201ϕ2-1	YP_001956966.1	virion structural protein	449	29.05	98	3.00E-161	476
Pseudomonas phage ϕPA3	AEH03604.1	virion structural protein	437	29.8	98	7.00E-169	495
Erwinia phage Ea35-70	YP_009004937.1	hypothetical protein Ea357_144	437	24.67	98	2.00E-141	425
Erwinia phage ϕEaH1	YP_009010232.1	virion structural protein	433	28.6	98	2.00E-148	442
Ralstonia phage RSL2	BAQ02585.1	hypothetical protein	433	24.28	98	8.00E-143	428
Vibrio phage JM-2012	YP_006383341.1	hypothetical protein TSMG0061	434	23.11	91	6.00E-89	289
Vibrio phage VP4B	AGB07322.1	hypothetical protein	424	24.19	93	1.00E-120	371
*Pseudomona*s phage OBP	YP_004958166.1	unnamed protein product	443	23.54	92	1.00E-132	403
Pseudomonas phage EL	YP_418202.1	hypothetical protein PPEV_gp169	448	23.62	97	5.00E-126	386

Table C.8. PSIBLAST hits for similar proteins of SPN3US gp214 in the ϕKZ-related phages. - refers to no match identified

Phage	Accession number	Title	aa length	% id	Query coverage	E-value	Bitscore
Erwinia phage ϕEaH2	YP_007237787.1	hypothetical protein	251	78.09	100	5.00E-115	342
Cronobacter phage CR5	YP_008125930.1	hypothetical protein CR5_190	253	47.22	100	4.00E-103	312
Pseudomonas phage ϕKZ	NP_803719.1	ORF153	304	19.72	82	3.00E-63	211
Pseudomonas phage 201ϕ2-1	YP_001956961.1	virion structural protein	316	14.08	82	1.00E-34	137
Pseudomonas phage ϕPA3	AEH03599.1	virion structural protein	320	14.35	83	3.00E-67	222
Erwinia phage Ea35-70							
Erwinia phage ϕEaH1	YP_009010228.1	hypothetical protein	285	19.53	94	0.004	47.9
Ralstonia phage RSL2	BAQ02579.1	hypothetical protein	293	13.27	81	7.00E-06	56
Vibrio phage JM-2012	YP_006383344.1	hypothetical protein TSMG0064	278	17.44	97	2.00E-72	234
Vibrio phage VP4B	-	-	-	-	-	-	-
*Pseudomona*s phage OBP	-	-	-	-	-	-	-
Pseudomonas phage EL	-	-	-	-	-	-	-

Table C.9. PSIBLAST hits for similar proteins of SPN3US gp219 in the ϕKZ-related phages

Phage	Accession number	Title	aa length	% id	Query coverage	E-value	Bitscore
Erwinia phage ϕEaH2	YP_007237791.1	putative virion structural protein	243	64.5	95	6.00E-87	270
Cronobacter phage CR5	YP_008125934.1	hypothetical protein CR5_194	240	44.05	94	1.00E-75	241
Pseudomonas phage ϕKZ	NP_803713.1	ORF147	238	23.87	95	1.00E-62	208
Pseudomonas phage 201ϕ2-1	YP_001956954.1	hypothetical protein 201phi2-1p231	241	22.5	94	2.00E-61	204
Pseudomonas phage ϕPA3	AEH03593.1	hypothetical protein	237	22.31	94	6.00E-60	200
Erwinia phage Ea35-70	YP_009004948.1	hypothetical protein Ea357_155	242	22.03	92	8.00E-59	198
Erwinia phage ϕEaH1	YP_009010222.1	hypothetical protein	236	17.03	92	5.00E-55	188
Ralstonia phage RSL2	BAQ02574.1	hypothetical protein	264	22.22	94	5.00E-66	217
Vibrio phage JM-2012	YP_006383348.1	hypothetical protein TSMG0068	238	22.03	93	3.00E-45	162
Vibrio phage VP4B	AGB07121.1	hypothetical protein	245	17.55	96	3.00E-41	152
*Pseudomona*s phage OBP	YP_004958177.1	unnamed protein product	244	17.92	80	1.00E-37	142
Pseudomonas phage EL	YP_418210.1	hypothetical protein PPEV_gp177	260	17.89	76	9.00E-35	136

Table C.10. PSIBLAST hits for similar proteins of SPN3US gp220 in the ϕKZ-related phages. - refers to no match identified

Phage	Accession number	Protein title	aa length	% id.	Query Coverage	E-value	Bitscore
Erwinia phage ϕEaH2							
Cronobacter phage CR5	YP_008125935.1	hypothetical protein CR5_195	155	39.19	94	1.00E-53	167
Pseudomonas phage ϕKZ	NP_803735.1	ORF169	171	21.9	66	2.00E-27	99.8
Pseudomonas phage 201ϕ2-1	YP_001956978.1	hypothetical protein 201phi2-1p255	175	20.62	60	6.00E-26	96
Pseudomonas phage ϕPA3	AEH03621.1	hypothetical protein	172	13.25	51	2.00E-20	81.7
Erwinia phage Ea35-70							
Erwinia phage ϕEaH1	YP_009010221.1	hypothetical protein	182	26.88	60	2.00E-19	79
Ralstonia phage RSL2	BAQ02573.1	hypothetical protein	99	24.19	40	0.06	28.6
Vibrio phage JM-2012	-	-	-	-	-	-	-
Vibrio phage VP4B	-	-	-	-	-	-	-
*Pseudomona*s phage OBP	-	-	-	-	-	-	-
Pseudomonas phage EL	-	-	-	-	-	-	-

Table C.11. PSIBLAST hits for similar proteins of SPN3US gp238 in the ϕKZ-related phages

Phage	Accession number	Title	aa length	% id	Query coverage	E-value	Bitscore
Erwinia phage ϕEaH2	YP_007237808.1	hypothetical protein	736	75.03	100	0	618
Cronobacter phage CR5	YP_008125946.1	hypothetical protein CR5_206	730	53.22	99	0	557
Pseudomonas phage ϕKZ	NP_803748.1	ORF182	664	22.49	99	8.00E-114	371
	NP_803748.1	ORF182		19.46	99	1.00E-110	362
Pseudomonas phage 201ϕ2-1	YP_001956999.1	hypothetical protein 201phi2-1p277	666	19.88	99	4.00E-104	345
	YP_001956999.1	hypothetical protein 201phi2-1p277		21.43	99	3.00E-112	366
Pseudomonas phage ϕPA3	AEH03637.1	hypothetical protein	658	19.79	99	2.00E-103	343
	AEH03637.1	hypothetical protein		20.9	99	1.00E-123	396
Erwinia phage Ea35-70	YP_009004953.1	hypothetical protein Ea357_160	662	22.4	99	7.00E-146	454
Erwinia phage ϕEaH1	YP_009010201.1	hypothetical protein	727	23.6	99	7.00E-149	464
Ralstonia phage RSL2	BAQ02568.1	hypothetical protein	631	25.39	99	4.00E-115	373
	BAQ02568.1	hypothetical protein		18.16	99	4.00E-88	301
Vibrio phage JM-2012	YP_006383354.1	hypothetical protein TSMG0074	699	20.33	97	2.00E-108	358
Vibrio phage VP4B	AGB07127.1	hypothetical protein	709	17.31	95	8.00E-137	432
Pseudomonas phage OBP	YP_004958182.1	unnamed protein product	733	19.53	99	4.00E-138	436
Pseudomonas phage EL	YP_418215.1	hypothetical protein PPEV_gp182	730	18.47	99	1.00E-136	432

Table C.12. PSIBLAST hits for similar proteins of SPN3US gp241 in the ϕKZ-related phages

Phage	Accession number	Title	aa length	% id	Query coverage	E-value	Bitscore
Erwinia phage ϕEaH2	YP_007237811.1	putative phage DNA-directed RNA polymerase beta subunit 2	1388	78.29	99	0	1497
Cronobacter phage CR5	YP_008125949.1	putative DNA-directed RNA polymerase beta subunit 3	1385	58.84	99	0	1416
Pseudomonas phage ϕKZ	NP_803744.1	ORF178	1451	28.78	99	0	1373
Pseudomonas phage 201ϕ2-1	YP_001956996.1	putative RNA polymerase beta subunit	1500	18.16	99	9.00E-80	300
	YP_001956996.1	putative RNA polymerase beta subunit		32.46	99	0	1081
Pseudomonas phage ϕPA3	AEH03632.1	hypothetical protein	379	38.87	27	1.00E-168	521
	AEH03634.1	putative RNA polymerase beta subunit	1097	24.15	73	0	988
Erwinia phage Ea35-70	YP_009004956.1	hypothetical protein Ea357_163	1528	24.15	99	3.00E-48	199
	YP_009004956.1	hypothetical protein Ea357_163		28.52	99	0	1127
Erwinia phage ϕEaH1	YP_009010198.1	RNA polymerase beta subunit	1504	21.05	99	2.00E-26	128
	YP_009010198.1	RNA polymerase beta subunit		27.03	99	0	1133
Ralstonia phage RSL2	BAQ02565.1	hypothetical protein	1472	26.63	98	0	1221
Vibrio phage JM-2012	YP_006383356.1	hypothetical protein TSMG0076	1293	24.71	97	0	1083
Vibrio phage VP4B	AGB07131.1	hypothetical protein	1451	24.88	99	0	1214
Pseudomonas phage OBP	YP_004958185.1	unnamed protein product	1450	25.35	98	0	1178
Pseudomonas phage EL	YP_418219.1	hypothetical protein PPEV_gp186	1068	23.48	73	0	852
	YP_418220.1	hypothetical protein PPEV_gp187	357	28.77	26	2.00E-134	429

Table C.13. PSIBLAST hits for similar proteins of SPN3US gp255 in the ϕKZ-related phages

Phage	Accession number	Title	aa length	% id	Query coverage	E-value	Bitscore
Erwinia phage ϕEaH2	YP_007237823.1	putative virion structural protein	292	90.38	100	3.00E-138	393
Cronobacter phage CR5	YP_008125959.1	putative virion structural protein 19	292	71.72	99	4.00E-129	370
Pseudomonas phage ϕKZ	NP_803596.1	ORF030	293	26.28	99	2.00E-115	335
Pseudomonas phage 201ϕ2-1	YP_001956757.1	major virion structural protein	291	24.91	99	5.00E-118	342
Pseudomonas phage ϕPA3	AEH03438.1	virion structural protein	291	26.28	99	1.00E-116	338
Erwinia phage Ea35-70	YP_009004970.1	tail tube protein	291	27.41	92	1.00E-102	302
Erwinia phage ϕEaH1	YP_009010129.1	putative virion structural protein	298	29.35	99	7.00E-107	314
Ralstonia phage RSL2	BAQ02556.1	hypothetical protein	288	31.08	99	8.00E-113	328
Vibrio phage JM-2012	YP_006383392.1	hypothetical protein TSMG0112	283	21.25	98	1.00E-86	261
Vibrio phage VP4B	AGB07158.1	hypothetical protein	298	20.16	80	2.00E-66	209
Pseudomonas phage OBP	YP_004957908.1	unnamed protein product	300	17.87	89	1.00E-65	208
Pseudomonas phage EL	YP_418038.1	hypothetical protein PPEV_gp005	302	19.77	90	3.00E-87	263

Appendix D. Secondary structures for homologs of SPN3US gp168.

Figure D.1. Secondary structure of gp168 of SPN3US phage (http://bioinf.cs.ucl.ac.uk/psipred/)

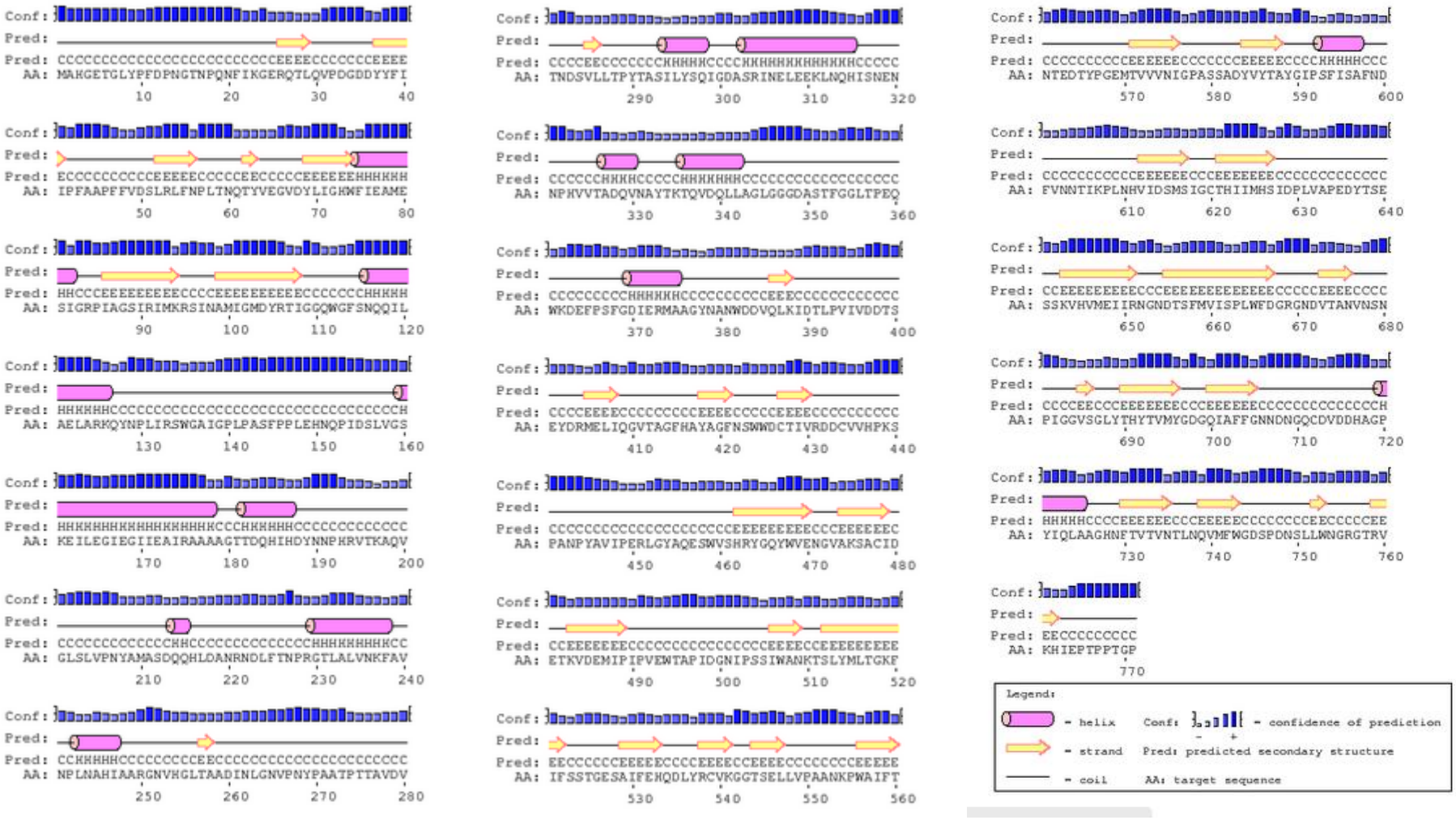

Figure D.2 Secondary structure of gp131 of ϕKZ phage (http://bioinf.cs.ucl.ac.uk/psipred/)

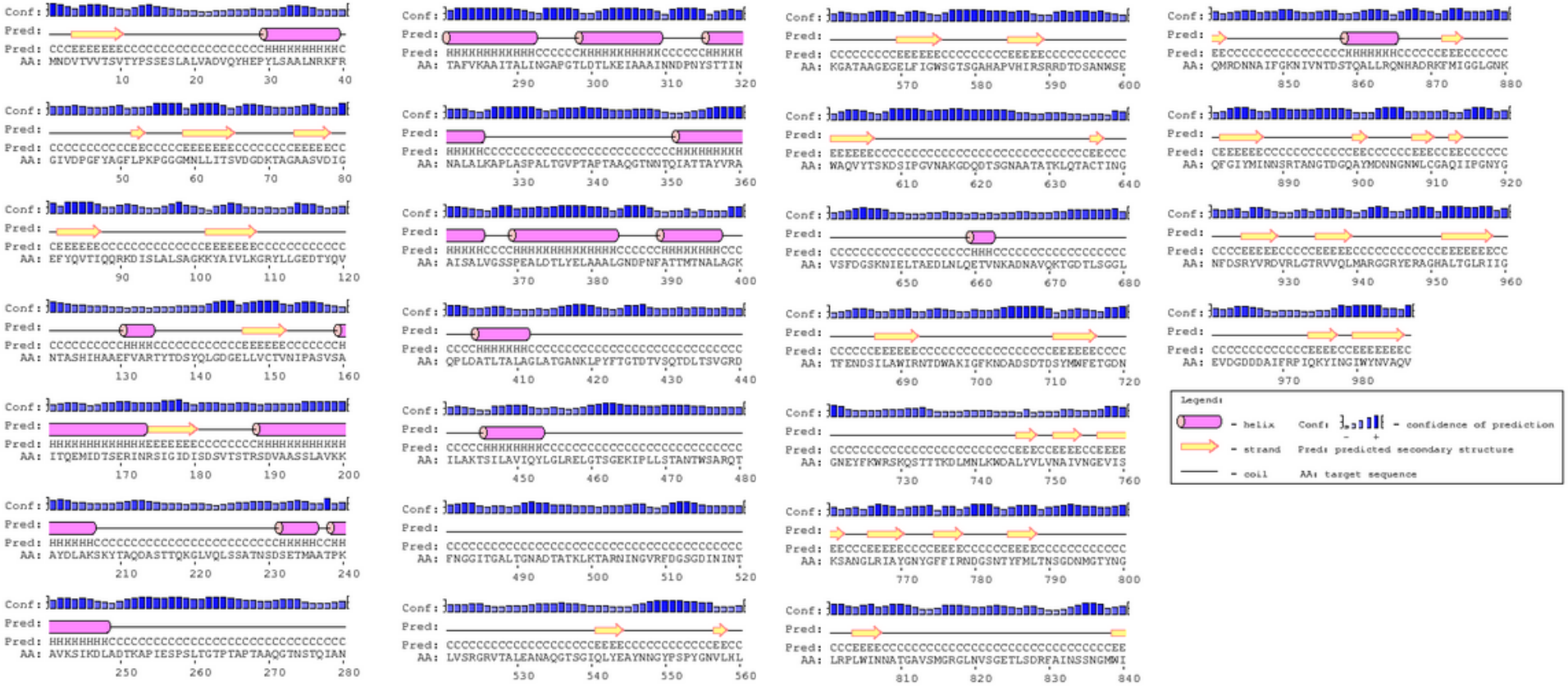

Figure D.3 Secondary structure of gpS of P1 phage (http://bioinf.cs.ucl.ac.uk/psipred/)

Appendix E. HHPRED results for SPN3US gp168 and ϕKZ gp131

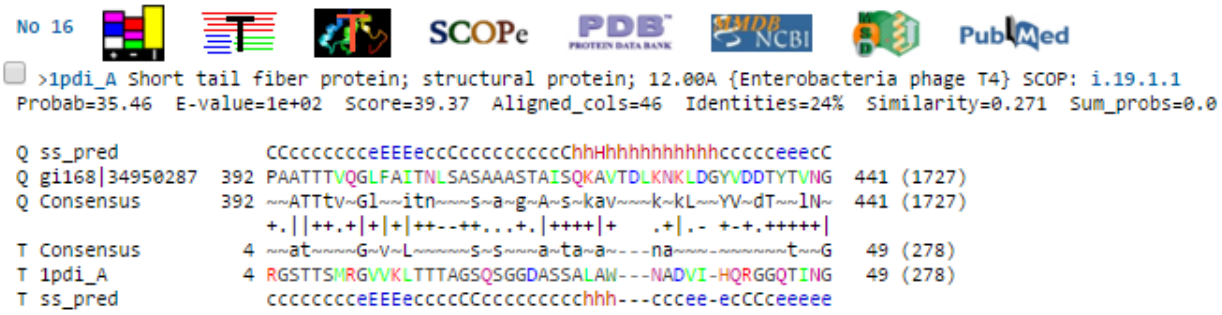

Figure E.1. HHPRED detection of homology of the whole gp168 protein in SPN3US (http://toolkit.tuebingen.mpg.de/hhpred)

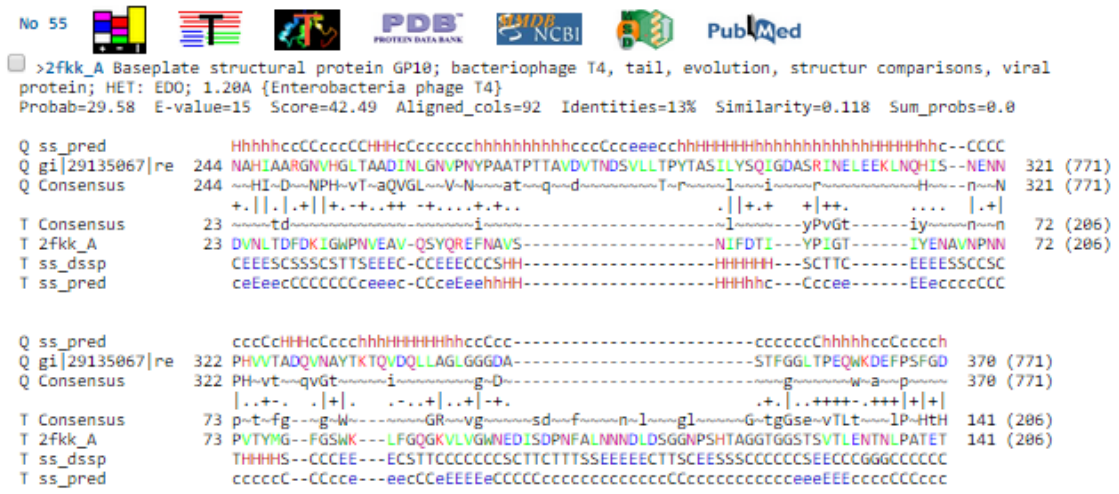

Figure E.2. HHPRED detection of homology of the whole gp131 protein in ϕKZ (http://toolkit.tuebingen.mpg.de/hhpred)

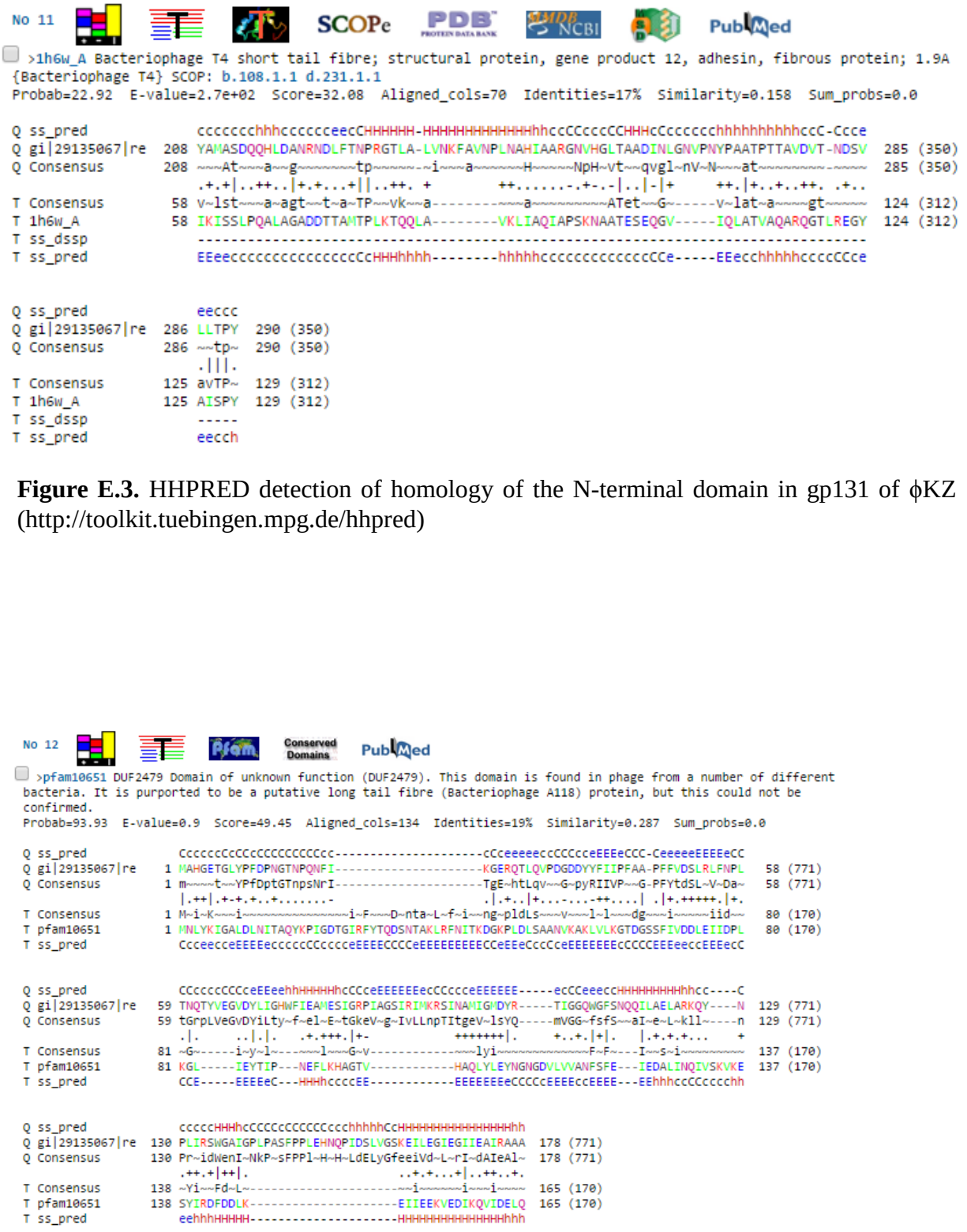

Figure E.3. HHPRED detection of homology of the N-terminal domain in gp131 of ϕKZ (http://toolkit.tuebingen.mpg.de/hhpred)

Figure E.4. HHPRED detection of homology of gp131 N-terminal domain in ϕKZ with a putative long tail fiber in phage A118. http://toolkit.tuebingen.mpg.de/hhpred